SYSTAT®

Statistics

SYSTAT: The System for Statistics

Copyright

Copyright © 1989, SYSTAT, Inc.: all rights reserved worldwide. No part of this publication may be reproduced, transmitted, transcribed, stored in a retrieval system, or translated into any language or computer language, in any form or by any means, electronic, mechanical, magnetic, optical, chemical, manual or otherwise, without the express written permission of SYSTAT, Inc.

References

SAS is a trademark of SAS Institute, Inc.
SYSTAT and SYGRAPH are registered trademarks of SYSTAT, Inc.

SYSTAT, Inc.
1800 Sherman Avenue
Evanston, IL 60201-3793
708.864.5670 Tel.
708.492.3567 FAX

Citation

The proper citation for this manual is:

Wilkinson, Leland. *SYSTAT: The System for Statistics.* Evanston, IL: SYSTAT, Inc., 1989.

Contents

1 Introduction to statistics 3

2 Cluster 20

3 Corr 50

4 Factor 72

5 MDS 94

6 MGLH 118
Regression 142
Analysis of variance 180
Multivariate models 282

7 Nonlin 308

8 Npar 354

9 Series 376

10 Stats 446

11 Tables 472

Appendix I: Command reference 506

Appendix II: Command equivalents for examples 553

Appendix III: Data files 575

Appendix IV: References 583

Index 599

Detailed tables of contents for each chapter follow.

Introduction to statistics 1

Descriptive statistics 3
Know thy batch 4
Sum, mean, and standard deviation 5
Stem-and-leaf diagrams 5
The median 6
Sorting 7
Standardizing 8
Inferential statistics 9
What is a population? 10
Picking a simple random sample 10
Specifying a model 11
Estimating a model 12
Confidence intervals 13
Hypothesis testing 14
Checking assumptions 16

© 1989, SYSTAT, Inc.

Cluster 2

Introduction 20
Types of clustering 20
Correlations and distances 21
Standardizing data 22
Hierarchical clustering 23
Clustering rows 23
Clustering columns 24
Clustering rows and columns 24
Partitioning via Kmeans... 25

Usage 27
Join... 27
Choosing distance metrics 28
Joining rows 29
Joining columns 29
Joining rows and columns 29
Labeling cases 30
Choosing linkage (amalgamation) methods 30
Saving files 31
Cluster displays 32
Kmeans... 34
Number of iterations 34
Number of groups 35
Labeling cases 35
Saving cluster membership 35

Computation 36
Memory limitations 36
Missing data 36
Algorithms 36

Examples 38
2.1 Joining with Euclidean distances 38
2.2 Joining with correlation metric 38
2.3 Joining a matrix 40
2.4 K-means 42
2.5 Distance matrix input 44

© 1989, SYSTAT, Inc.

Corr 3

Introduction 50
The scatterplot matrix (SPLOM) 50
The Pearson correlation coefficient 51
Other measures of association 52
Measures for rank order data 52
Measures for binary data 53
Transposed data 55
Pairwise and listwise deletion of missing data 56
Usage 58
City... 58
Covariance... 59
Dichotomous... 59
Euclidean... 60
Gamma... 60
Mu2...
Pearson... 61
Probabilities for Pearson correlation coefficients 62
SSCP... 62
Spearman... 62
Tau... 63
Saving files 63
Computation 64
Missing data 64
Algorithms 64
Examples 65
3.1 Pearson correlation 65
3.2 Probabilities 66
3.3 Bonferroni-adjusted probabilities 67

© 1989, SYSTAT, Inc.

Factor 4

Introduction 72
A principal component 72
Multiple principal components 74
Exploratory vs. confirmatory components 75
Component coefficients 76
Component loadings 76
Correlations or covariances? 76
Rotation 77
Signs of component loadings 77
Factor analysis 78
Estimating factors 79
Principal components vs. factor analysis 79
Applications and caveats 82
Usage 83
Principal components... 83
Type of matrix to factor 84
Pairwise deletion of missing data 84
Number of factors and Eigenvalue 85
Rotation 85
Plot factors 85
Labels 85
Sort loadings 85
Save file 85
Show complete results 86
Common factor model... 86
Iterations 86
Tolerance 86
Save file 86
Computation 87
Missing data 87
Algorithms 87
Examples 88
4.1 Principal components 88
4.2 Plot component scores 90

© 1989, SYSTAT, Inc.

MDS 5

Introduction 94
Assumptions 94
Collecting dissimilarity data 95
Direct methods 95
Indirect methods 96
Scaling dissimilarities 97
Stress and iterations 97
The Shepard diagram 97
The **MDS** plot 97
An example 98

Usage 100
Scale... 100
Controlling convergence 101
Number of dimensions 101
Choosing a distance metric 102
Shepard plots 102
Choosing a loss function 102
Choosing a regression function 102
Saving files 102
Configuration... 103
Confirmatory scaling 104

Computation 105
Missing data 105
Algorithms 105
Kruskal method 105
Guttman method 106

Examples 108
5.1 Guttman method 108
5.2 Kruskal method 111

© 1989, SYSTAT, Inc.

MGLH 6

Introduction 118
Simple linear models 118
Regression 120
Least squares 122
Estimation and inference 122
Standard errors 124
Hypothesis testing 124
The multiple correlation 126
Regression diagnostics 126
Multiple regression 130
Variable selection 132
Multivariate regression 134
Matrix notation 134
The multivariate general linear hypothesis 135
Analysis of variance 136
Two-group ANOVA 136
Multi-group ANOVA 138
Multivariate ANOVA 139
Conclusion 139

Computation 140
Missing data 116
Algorithms 116

© 1989, SYSTAT, Inc.

Regression

Introduction 143

Usage 145
Regression... 146
Saving residuals 146
Regression without the constant 146
General linear model... 147
Previewing your model 148
Saving residuals and other results 148
Cross validation 149
Durbin-Watson statistic 150
Regression without the constant 150
Repeated measures 150
Stepwise regression 150
Mixture models 150
Using an SSCP, covariance, or correlation matrix as input 150
Test... 152
Options 154
Selecting coefficients 155
Weighting coefficients 155
Show complete results 157

Examples 158
6.1 Simple regression 158
Saving residuals 160
6.2 Multiple regression 160
Extended output 162
Saving residuals 164
6.3 Testing hypotheses 165
Testing one regression coefficient 165
Testing two regression coefficients 166
Testing the same two coefficients with an A matrix 167
Testing a nonzero null hypothesis with a D matrix 168
6.4 Regression with ecological or grouped data 168
6.5 Weighted least squares 168
6.6 Regression without the constant 172
6.7 Mixture models 173
6.8 Polynomial regression 175
6.9 Stepwise regression 176
Options to control stepping 177
6.10 Partial correlations 178

© 1989, SYSTAT, Inc.

Analysis of variance

Introduction 181

Usage 182
Fully factorial (M)ANOVA... 183
Saving residuals 184
Repeated measures 184
General linear model... 184
Options 185
Previewing your model 186
Repeated measures 186
Saving residuals and other results 186
Durbin-Watson statistic 188
ANOVA without the constant 188
Test... 188
Options 190
Selecting effects for hypothesis tests: 190
Comparisons and contrasts 191
Selecting error terms: 193
[Profile] 194
Show complete results 194

Examples 195
6.11 Analysis of variance 195
6.12 Two-way analysis of variance 202
6.13 Single degree of freedom designs 212
6.14 Unbalanced designs 216
6.15 Randomized block designs 219
6.16 Incomplete block designs 220
6.17 Fractional factorial designs 221
6.18 Mixed models 223
6.19 Nested designs 225
6.20 Split plot designs 227
6.21 Latin square designs 230
6.22 Crossover and changeover designs 232
6.23 Analysis of covariance 234
6.24 Hotelling's T^2 237
6.25 Multivariate analysis of variance 238
6.26 Multivariate analysis of covariance 244
6.27 Multivariate profile analysis 245
6.28 Unweighting means 250
6.29 One-way repeated measures 254
6.30 Repeated measures with grouping factors 256
6.31 Repeated measures with two trial factors 264
6.32 Repeated measures with three or more trials factors 270
6.33 Repeated measures analysis of covariance 274
6.34 Covariance alternatives to repeated measures 277
6.35 Issues in repeated measures analysis 277

© 1989, SYSTAT, Inc.

Multivariate models

Introduction 283

Usage 284
General linear model... 284
Options 285
Using an SSCP, covariance, or correlation matrix as input 285
Previewing your model 286
Saving residuals and other results 287
Cross validation 287
Regression without the constant 287
Test... 287
Options 288
Selecting effects for hypothesis tests 288
Comparisons and contrasts 288
Choosing prior probabilities 288
Standardizing canonical coefficients 290
Factoring the hypothesis or error matrix 290
Saving results 291
Show complete results 291

Examples 292
6.36 Discriminant analysis 292
Saving canonical scores 296
Classification tables 297
Canonical variable plot 298
Prior probabilities 298
Cross validation 298
Stepwise discriminant analysis 299
6.37 Within groups principal components analysis 299
6.38 Canonical correlation analysis 304

© 1989, SYSTAT, Inc.

Nonlin 7

Introduction 308
Modeling the dose-response function 308
Loss functions 312
Maximum likelihood 312
Least absolute values 312

Usage 313
Model... 313
Estimation method 316
Initial values 316
Iterations 316
Convergence 316
Maximum likelihood estimation 317
Printing standard errors 317
Saving files 317
Loss function... 317
Resume 319
Show complete results 319
Operation 319
Asymptotic standard errors and parameter intercorrelations 319
Parameter bounds and model constraints 320
Problems 321

Computation 323
Missing data 323
Algorithms 323

Examples 325
7.1 Estimating a mean 325
7.2 Estimating a median 326
7.3 Linear regression 327
7.4 Least absolute values regression 328
7.5 Nonlinear regression 328
7.6 Piecewise regression 333
7.7 Maximum likelihood estimation 337
7.8 Logistic models—iteratively reweighted least squares 339
7.9 Logistic models—maximum likelihood estimation 340
7.10 Probit models—maximum likelihood estimation 342
7.11 Kinetic models 345
7.12 Deterministic models—minimizing an analytic function 348
7.13 A random model 349
7.14 Contouring the loss function 349

© 1989, SYSTAT, Inc.

Npar 8

Introduction 354
Rank (ordinal) data 354
Categorical (nominal) data 354
Robustness 355
An example 355

Usage 357
Friedman... 357
Kolmogorov-Smirnov... 358
One-sample tests 359
Two-sample tests 359
Lilliefors test 359
Kruskal-Wallis... 360
Sign test... 360
Wald-Wolfowitz... 361
Wilcoxon... 361

Computation 362
Missing data and ties 362
Algorithms 362

Examples 363
8.1 Sign test 363
8.2 Wilcoxon signed ranks test 364
8.3 Friedman two-way analysis of variance 365
8.4 Kruskal-Wallis one-way analysis of variance 365
8.5 Wald-Wolfowitz runs test 367
8.6 Kolmogorov-Smirnov one-sample test 368
8.7 Lilliefors test 370
8.8 Kolmogorov-Smirnov two-sample test 371

© 1989, SYSTAT, Inc.

Series 9

Introduction 376
Smoothing 376
Moving averages 377
Weighted running smoothing 380
Running median smoothers 380
LOWESS 381
And the winner is... 382
ARIMA modeling and forecasting 383
Autocorrelation 385
Autoregressive models 386
Moving average models 388
ARMA models 388
Identifying models 389
Estimating the ARIMA model 392
Problems with forecast models 392
Seasonal decomposition and adjustment 393
Exponential smoothing 394
Fourier analysis 395

Usage 397
Transform... 398
ID... 401
Clear 401
Case plot... 402
ACF plot... 403
PACF plot... 403
CCF plot... 404
ARIMA... 404
Smooth... 407
LOWESS... 407
Exponential... 408
Seasonal... 410
Fourier... 412

Computation 414
Missing data 414
Algorithms 414

Examples 416
9.1 Series plots and differencing 416
9.2 ACF, PACF, CCF plots 419
9.3 ARIMA model identification 425
9.4 ARIMA with backcasting 430
9.5 ARIMA with forecasting 433
9.6 Exponential smoothing 434
9.7 Seasonal decomposition and adjustment 437
9.8 Fourier analysis 439

© 1989, SYSTAT, Inc.

Stats 10

Introduction 446
Location 446
Spread 447
The normal distribution 447
Nonnormal shape 448
Subpopulations 450
One-way ANOVA and t-tests 451

Usage 453
Statistics... 453
Statistics by group 454
Saving statistics 454
t-test... 455
Independent *t*-tests 456
Paired *t*-tests 456
Matched pairs 456
One-sample *t*-tests 457
One-way ANOVA... 457
Post hoc pairwise comparisons 458

Computation 459
Missing data 459
Algorithms 459

Examples 460
10.1 Basic statistics 460
10.2 Statistics by groups 460
10.3 Two sample *t*-test 462
10.4 Dependent *t*-test 463
10.5 Analysis of variance 464

© 1989, SYSTAT, Inc.

Tables 11

Introduction 472
Making tables 472
One-way tables 472
Two-way tables 473
Standardizing tables with percents 473
Significance tests 475
One-way tables 475
Two-way tables 477
McNemar's test 480
Multi-way tables 480
Measures of association for two-way tables 480
Phi, Cramer's V, and contingency 481
Cohen's kappa 481
Ordinal association 481
Measures for 2×2 tables 481
Log-linear models 481

Usage 483
Tabulate... 483
One-way tables 484
Two-way tables 484
Multi-way tables 484
Options 485
Tests and measures of association for two-way tables 486
Large tables 486
Log-linear models... 487
Options 489
Saving tables 490

Computation 491
Missing data 491
Algorithms 491

Examples 492
11.1 One-way frequency table 492
11.2 List table 492
11.3 Confidence intervals 493
11.4 Two-way table with tests and association measures 494
11.5 Percents 496
11.6 Mantel-Haenszel test 497
11.7 Log-linear model 499
11.8 Log-linear model of multi-way tables 500

© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

1 Introduction to statistics

Descriptive statistics	3
Know thy batch	4
Sum, mean, and standard deviation	5
Stem-and-leaf diagrams	5
The median	6
Sorting	7
Standardizing	8
Inferential statistics	9
What is a population?	10
Picking a simple random sample	10
Specifying a model	11
Estimating a model	12
Confidence intervals	13
Hypothesis testing	14
Checking assumptions	16

© 1989, SYSTAT, Inc.

Statistics and *state* have the same root. Statistics are the numbers of the state. More generally, they are any numbers or symbols that formally summarize our observations of the world. As we all know, summaries can mislead or elucidate. Statistics also refers to the introductory course we all seem to hate in college. When taught well, however, it is this course that teaches us how to use numbers to elucidate rather than to mislead.

Statisticians specialize in many areas—probability, exploratory data analysis, modeling, social policy, decision making, and others. While they may philosophically disagree, statisticians nevertheless recognize at least two fundamental tasks: *description* and *inference*. Description involves characterizing a batch of data in simple but informative ways. Inference involves generalizing from a sample of data to a larger population of possible data. Descriptive statistics help us to observe more acutely, and inferential statistics help us to formulate and test hypotheses.

Any distinctions, such as this one between descriptive and inferential statistics, are potentially misleading. Let's look at some examples, however, to see some differences between these approaches.

Descriptive statistics

Descriptive statistics may be single numerical summaries of a batch, such as an average. Or, they may be more complex tables and graphs. What distinguishes descriptive statistics is their reference to a given batch of data rather than to a more general population or class. While there are exceptions, we usually examine descriptive statistics to understand the structure of a batch. A closely related field is called *exploratory data analysis*. Both exploratory and descriptive methods may lead us to formulate laws or test hypotheses, but their focus is on the data at hand.

© 1989, SYSTAT, Inc.

Consider, for example, the following batch. These are numbers of arrests by sex in 1985 for selected U.S. crimes. The source is the FBI Uniform Crime Reports. What can we say about differences between the patterns of arrests of U.S. men and women in 1985?

CRIME	MALES	FEMALES
murder	12904	1815
rape	28865	303
robbery	105401	8639
assault	211228	32926
burglary	326959	26753
larceny	744423	334053
auto	97835	10093
arson	13129	2003
battery	416735	75937
forgery	46286	23181
fraud	151773	111825
embezzle	5624	3184
vandal	181600	20192
weapons	134210	10970
vice	29584	67592
sex	74602	6108
drugs	562754	90038
gambling	21995	3879
family	35553	5086
dui	1208416	157131
drunk	726214	70573
disorderly	435198	99252
vagrancy	24592	3001
runaway	53808	72473

Know thy batch

First, we must be careful in characterizing the batch. These statistics do not cover the gamut of U.S. crimes. We left out Curfew and Loitering violations, for example. Not all *reported* crimes are included in these statistics. Some false arrests may be included. State laws vary on the definitions of some of these crimes. Agencies may modify arrest statistics for political purposes. Some transvestite criminals may have been misidentified. And so on. This begins to sound like a warranty disclaimer! Know where your batch came from before you use it.

© 1989, SYSTAT, Inc.

Sum, mean, and standard deviation

Were there more male than female arrests for these crimes in 1985? The following **Stats** output shows us the answer. Males were arrested for 5,649,688 crimes (*not* 5,649,688 males—some may have been arrested more than once). Females were arrested 1,237,007 times.

```
Total observations:   24

                        MALE        FEMALE

N of cases                24            24
Minimum               5624.0         303.0
Maximum            1208416.0      334053.0
Mean                235403.7       51542.0
Standard dev        305947.1       74220.9
Sum                5649688.0     1237007.0
```

How about the average (*mean*) number of arrests for a crime? For males, this was 235,403 and for females, 51,542. Does the mean make any sense to you as a summary statistic? Another statistic in the table, the *standard deviation*, measures how much these numbers vary around the average. The standard deviation is the square root of the average squared deviation of the observations from their mean. It, too, has problems in this instance. First of all, both the mean and standard deviation should represent what you *could* observe in your batch, on average: the mean number of fish in a pond, the mean number of children in a classroom, the mean number of red blood cells per cubic millimeter. Here, we would have to say, "the mean *murder-rape-robbery-...-runaway* type crime." Second, even if the mean made sense descriptively, we might question its use as a typical crime-arrest statistic. To see why, we need to examine the *shape* of these numbers.

Stem-and-leaf diagrams

Let's look at a display that compresses these data a little less drastically. The stem-and-leaf diagram is like a tally. We pick a most significant digit or digits and tally the next digit to the right. By using trailing digits instead of tally marks, we preserve extra digits in the data. You can read more about this display in the graphics manual, but for now, notice the shape of the tally. There are mostly smaller numbers of arrests, and a few crimes (e.g. driving under the influence of alcohol and larceny) with larger numbers of arrests. Another way of saying this is that it is *positively skewed* toward larger numbers for both males and females.

© 1989, SYSTAT, Inc.

```
        Stem and leaf plot of variable:     MALE     , N =    24

Minimum:          5624.000
Lower hinge:     29224.500
Median:         101618.000
Upper hinge:    371847.000
Maximum:       1208416.000

                0 H 011222234579
                1 M 0358
                2   1
                3 H 2
                4   13
                5   6
                6
                7   24
         ***Outside values***
              1   0


        Stem and leaf plot of variable:   FEMALE     , N =    24

Minimum:           303.000
Lower hinge:      4482.500
Median:          21686.500
Upper hinge:     74205.000
Maximum:        334053.000

                0   00000000011
                0 M 2223
                0
                0 H 6777
                0   99
                1   1
                1
                1   5
         ***Outside values***
                3   3
```

The median

When data are skewed like this, the mean gets pulled from the center of the majority of numbers toward the extreme with the few. A statistic that is not as sensitive to extreme values is the *median*. The median is the value above which half the data fall. More precisely, if you sort the data, the median is the middle value or the average of the two middle values. Notice that for males the median is 101,618, and for females, 21,686. Both are considerably smaller than the means and more typical of the majority of the numbers. This is why the median is often used for representing skewed data, such as incomes, populations, or reaction times.

We still have the same representativeness problem we had with the mean, however. Even if the medians corresponded to real data values in this batch (which they don't, since there is an even number of observations), it would be hard to characterize what they would represent.

© 1989, SYSTAT, Inc.

Sorting

Most people think of means, standard deviations, and medians as the primary descriptive statistics. They are useful summary quantities when the observations represent values of a single variable. We purposely chose an example where they are less appropriate, however, even when they are easily computable. There are better ways to reveal the patterns in these data. Let's look at *sorting* as a way of uncovering structure.

We were talking once with an FBI agent who had helped to uncover the Chicago machine's voting fraud scandal some years ago. He was a statistician, so we were curious what statistical methods he used to prove the fraud. He replied, "We sorted the voter registration tape alphabetically by last name. Then we looked for duplicate names and addresses." Sorting is one of the most basic and powerful data analysis techniques. The stem-and-leaf diagram, for example, is a sorted display.

We can sort on any numerical or character variable. It depends on our goal. We began this chapter with a question: are there differences between the patterns of arrests of U.S. men and women in 1985? How about sorting the male and female arrests separately? If we do this, we will get a list of crimes in order of decreasing frequency within sex.

MALES	FEMALES
dui	larceny
larceny	dui
drunk	fraud
drugs	disorderly
disorderly	drugs
battery	battery
burglary	runaway
assault	drunk
vandal	vice
fraud	assault
weapons	burglary
robbery	forgery
auto	vandal
sex	weapons
runaway	auto
forgery	robbery
family	sex
vice	family
rape	gambling
vagrancy	embezzle
gambling	vagrancy
arson	arson
murder	murder
embezzle	rape

© 1989, SYSTAT, Inc.

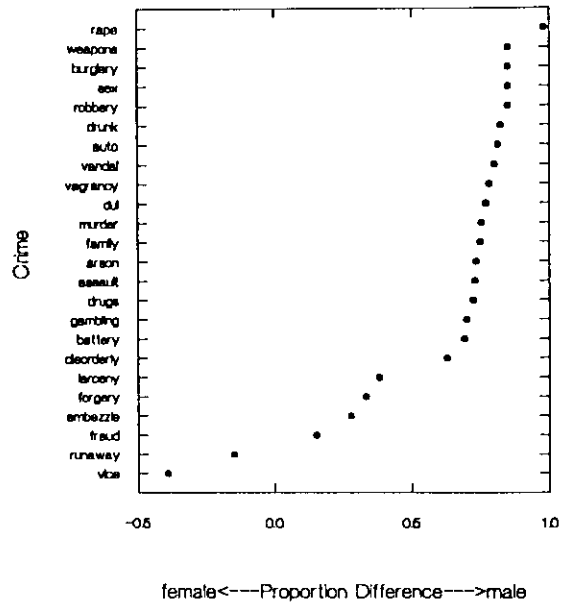
You might want to connect similar crimes with lines. The number of crossings would indicate differences in ranks.

Standardizing

This ranking is influenced by prevalence. The most frequent crimes occur at the top of the list in both groups. Comparisons within crimes are obscured by this influence. Men committed almost 100 times as many rapes as women, for example, yet rape is near the bottom of both lists. If we are interested in contrasting the sexes on patterns of crime while holding prevalence constant, we must *standardize* the data. There are several ways to do this. You may have heard of standardized test scores for aptitude tests. These are usually produced by subtracting means and then dividing by standard deviations. Another method is simply to divide by row or column totals. For the crime data, we will divide by totals within rows (each crime). Doing so gives us the proportion of each arresting crime committed by men or women. The total of these two proportions will thus be 1.

Now, a contrast between men and women on this standardized value should reveal variations in arrest patterns within crime type. By subtracting the female proportion from the male, we will highlight primarily male crimes with positive values and female crimes with negative. Next, sort these differences and plot them in a simple graph. Here is the result.

© 1989, SYSTAT, Inc.



Now we can see clear contrasts between males and females in arrest patterns. The predominantly aggressive crimes appear at the top of the list. Rape now appears where it belongs—an aggressive rather than sexual crime. A few crimes dominated by females are at the bottom.

Inferential statistics

We often want to do more than describe a particular sample. In order to generalize, formulate a policy, or test an *hypothesis*, we need to make an *inference*. Making an inference implies that we think a model describes a more general *population* from which our data have been randomly *sampled*. Sometimes it is difficult to imagine a population from which you have gathered data. A population can be "all possible voters," "all possible replications of this experiment," or "all possible moviegoers." When you make inferences, you should have a population in mind.

© 1989, SYSTAT, Inc.

What is a population?

We are going to use inferential methods to estimate the mean age of an unusual population. It is contained in a two volume book called *Who's Who in America* (1980). If we had infinite patience, we could enter all 73,500 ages into a SYSTAT file and compute the mean age exactly. This would be preferable if practical. Sometimes, however, a sampling estimate can be more accurate than an entire census. This occurs when biases are introduced in large censuses from refusals to comply, keypunch or coding errors, and other sources. In these cases, a carefully constructed random sample can yield less-biased information about the population.

We called this an unusual population because it is contained in a book and therefore finite. We are not about to estimate the mean age of the rich and famous. After all, *Spy* magazine has a regular feature listing all the famous people who are *not* in *Who's Who*. And college students everywhere send in fake listings that may escape the careful fact checking of the Marquis staff. When we get our estimate, we might be tempted to generalize beyond the book, but we would be wrong to do so. This is an important point. If a psychologist measures opinions in a random sample from a class of college sophomores, her conclusions should begin with the statement, "College sophomores at my university think..." If the word "people" is substituted for "college sophomore," it is the experimenter's responsibility to make clear that the sample is representative of the larger class on all attributes that might affect the results.

Picking a simple random sample

That our population is finite should cause us no problems as long as our sample is much smaller than the population. Otherwise, we would have to use special techniques to adjust for the bias it would cause. How do we choose a *simple random sample* from a population? We use a method that insures that every possible sample of a given size has an equal chance of being chosen. Here are some methods that are *not* random:

The first name on every tenth page (some names have no chance of being chosen).
Close your eyes, riffle the book, and point to a name (Tversky and others have done research that shows that humans cannot behave randomly).
Randomly pick the first letter of the last name and randomly choose from the names beginning with that letter (there are more names beginning with, say, C than with I).

© 1989, SYSTAT, Inc.

The way to pick randomly from a book, file, or any finite population, is to assign a number to each name or case and then pick a sample of numbers randomly. You can use SYSTAT to generate a random number between 1 and 73500, for example, with the expression:

```
1 + INT(73500*URN)
```

We did this, started to count names, and gave up. There were too many names to count. As a short cut, we randomly generated a page number and picked a name off the page using the random number generator. This method should work well *provided* each page has approximately the same number of names (between 19 and 21 in this case). Anyway, here is the sample:

AGE	SEX	*(continued)*	
60	male	38	female
74	male	44	male
39	female	49	male
78	male	62	male
66	male	76	female
63	male	51	male
45	male	51	male
56	male	75	male
65	male	65	female
51	male	41	male
52	male	67	male
59	male	50	male
67	male	55	male
48	male	45	male
36	female	49	male
34	female	58	male
68	male	47	male
50	male	55	male
51	male	67	male
47	male	58	male
81	male	76	male
56	male	70	male
49	male	69	male
58	male	46	male
58	male	60	male

Specifying a model

To make an *inference* about age we need to construct a model for our population.

$$a = \mu + \varepsilon$$

© 1989, SYSTAT, Inc.

This model says that the age (*a*) of someone we pick from the book can be described by an overall mean age (μ) plus an amount of error (ε) specific to that person and due to random factors that are too numerous and insignificant to describe systematically. Notice that we use Greek letters to denote things we cannot observe directly and Roman for those we do observe. Of the unobservables in the model, μ is called a ***parameter*** and ε a ***random variable***. A parameter is a constant that helps to describe a population. Parameters indicate how a model is an instance of a family of models for similar populations. A random variable varies like the tossing of a coin.

There are two more parameters associated with the random variable ε but not appearing in the model equation. One is its mean (μ_ε), which we have rigged to be 0, and its standard deviation (σ_ε or simply σ). Because *a* is simply the sum of μ (a constant) and ε (a random variable), its standard deviation is also σ.

In specifying this model, we usually assume:

1. It is true for every member of the population.
2. The error, plus or minus, that helps determine one population member's age is ***independent*** of (not predictable from) the error for other members.
3. The errors in predicting all the ages come from the same random distribution with mean zero and standard deviation σ.

Estimating a model

Because we have not sampled the entire population, we cannot compute the parameter values directly from the data. We have only a small sample from a much larger population, so we can only ***estimate*** the parameter values using some statistical method on our sample data. When our three assumptions are appropriate, the sample average will be a good estimate of the population mean. Without going into all the details, the sample estimate will be, on average, close to the values of the mean in the population.

We can use various methods in SYSTAT to estimate the mean. One way is to put our model directly to **MGLH**, selecting AGE as the [Dependent] variable and just clicking [Indep'dent] to get a CONSTANT term.

© 1989, SYSTAT, Inc.

This model says that age is a function of a constant value (μ). The rest is error (ε). Another method is to compute the mean from the basic Stats routines. Here is the result:

```
                    AGE

N OF CASES           50
MEAN              56.700
STANDARD DEV      11.620
STD. ERROR         1.643
```

Our best estimate of the mean age of people in *Who's Who* is 56.7 years.

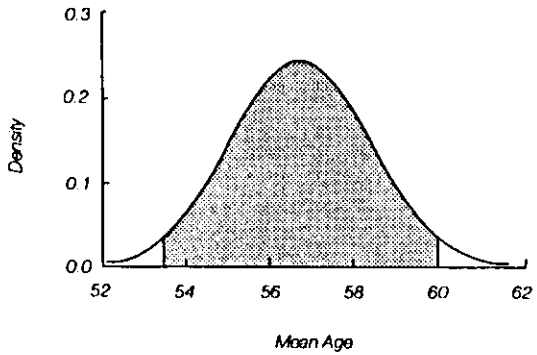
Confidence intervals

Our estimate seems reasonable, but it is not exactly correct. If we took more samples of size 50 and computed estimates, how much would we expect them to vary? First, it should be plain without any mathematics to see that the larger our sample, the closer will be our sample estimate to the true value of μ in the population. After all, if we could sample the entire population, the estimates would be the true values. Even so, the variation in sample estimates is a function only of the sample size and the variation of the ages in the population. It does not depend on the size of the population (number of people in the book). Specifically, the standard deviation of the sample mean is the standard deviation of the population divided by the ***square root*** of the sample size. This ***standard error of the mean*** is listed on the output above as 1.643. On average, we would expect our sample estimates of the mean age to vary by plus or minus a little more than one and a half years, assuming samples of size 50.

If we knew the ***shape*** of the sampling distribution of mean age, we would be able to complete our description of the accuracy of our estimate. There is an approximation that works quite well, however. If the sample size is reasonably large (say, greater than 25), then the mean of a simple random sample is approximately ***normally distributed***. This is true even if the population distribution is not normal, provided the sample size is large.

© 1989, SYSTAT, Inc.

We now have enough information from our sample to construct a normal approximation to the distribution of our sample mean. The following figure shows this approximation centered at the sample estimate of 56.7 years. Its standard deviation is taken from the standard error of the mean, 1.643 years.



We have drawn the graph so that the central shaded area comprises 95 percent of all the area under the curve (from about 53.4 to 59). From this normal approximation, we have built a 95 percent *symmetric confidence interval* that gives us a specific idea of the variability of our estimate. If we did this entire procedure again—sample 50 names, compute the mean and its standard error, and construct a 95 percent confidence interval using the normal approximation—then we would expect that 95 intervals out of a hundred so constructed would cover the real population mean age. Remember, population mean age is not necessarily at the center of the interval we just constructed, but we do expect the interval to be close to it.

Hypothesis testing

From the sample mean and its standard error, we can also construct *hypothesis tests* on the mean. Suppose someone believed the average age of *Who's Who*'ers is 61 years. After all, we might have picked an unusual sample just through luck of the draw. Let's say, for argument, that the population mean age *is* 61 and the standard deviation is 1.643. How likely would it be to find a sample mean age of 56.7? If it is very unlikely, then we would reject this *null hypothesis* that the population mean is 61. Otherwise, we would fail to reject it.

© 1989, SYSTAT, Inc.

There are several ways to represent an *alternative hypothesis* against this null hypothesis. We could make a *simple alternative* value 56.7 years. Usually, however, we make the alternative *composite*—that is, it represents a range of possibilities that do not include the value 61. Here is how it would look:

$$H_0: \mu = 61 \text{ (null hypothesis)}$$

$$H_A: \mu \neq 61 \text{ (alternative hypothesis)}$$

We would reject the null if our sample value for the mean were outside a set of values that a population value of 61 could plausibly generate. In this context, "plausible" means more probable than a conventionally agreed upon *critical level* for our test. This value is usually .05. A result that would be expected to occur fewer than five times in a hundred samples is considered *significant* and would be a basis for rejecting our null hypothesis.

Constructing this hypothesis test is mathematically equivalent to sliding the normal distribution in the above figure to center over 61. We then look at the sample value 56.7 to see if it is outside the shaded area. If so, we reject the null hypothesis. Here is MGLH output that shows a *p value* (probability) of .012 for this test. Because this value is lower than .05, we would reject the null hypothesis that the mean age is 61. This is equivalent to saying that the value of 61 does not appear in the 95 percent confidence interval.

TEST OF HYPOTHESIS

SOURCE	SS	DF	MS	F	P
HYPOTHESIS	924.500	1	924.500	6.847	0.012
ERROR	6616.500	49	135.031		

The mathematical duality between confidence intervals and hypothesis testing may lead you to wonder which is more useful. The answer is that it depends on the context. Scientific journals usually follow a hypothesis testing model because their null hypothesis value for an experiment is usually zero and the scientist is attempting to reject the hypothesis that nothing happened in the experiment. Any rejection is usually taken to be interesting, even when the sample size is so large that even tiny differences from zero will be detected.

© 1989, SYSTAT, Inc.

Those involved in making decisions—epidemiologists, business people, engineers—are often more interested in confidence intervals. They focus on the size and credibility of an effect and care less whether it can be distinguished from zero. Some statisticians, called *Bayesians*, go a step further and consider statistical decisions as a form of betting. They use sample information to modify prior hypotheses. While there is no room here to discuss these issues, you might want to read Box and Tiao (1973) or Berger (1985) for further information on Bayesian statistics.

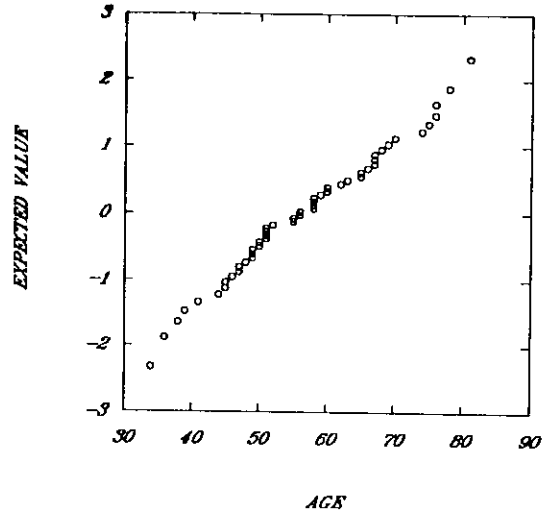
Checking assumptions

Now that we have finished our analyses, we should check some of the assumptions we made in doing them. First, we should examine whether the data look normally distributed. Although sample means will tend to be normally distributed even when the population isn't, it helps to have a normally distributed population, especially when we do not know the population standard deviation. The stem-and-leaf diagram gives us a quick idea:

```
        Stem and leaf plot of variable:      AGE     , N =    50

Minimum:              34.000
Lower hinge:          49.000
Median:               56.000
Upper hinge:          66.000
Maximum:              81.000

              3   4
              3   689
              4   14
              4 H 556778999
              5   0011112
              5 M 556688889
              6   0023
              6 H 55677789
              7   04
              7   5668
              8   1
```


© 1989, SYSTAT, Inc.

A better test of normality is to plot the sorted age values against the corresponding values of a mathematical normal distribution. This is called a *normal probability plot*. If the data are normally distributed, then the plotted values should fall approximately on a straight line. Our data plot fairly straight. The slight curvature at the ends indicate that the tails are shorter than we would expect from a normal distribution.



There is one more thing to notice. Take a look at the ages of the women in the sample. The four youngest members of the sample are all women. Does this invalidate our results? No, but it suggests that we might want to examine the gender differences further to see whether they are significant. Data analysis never ends.

© 1989, SYSTAT, Inc.

2 Cluster

Introduction 20
Types of clustering 20
Correlations and distances 21
Standardizing data 22
Hierarchical clustering 23
Clustering rows 23
Clustering columns 24
Clustering rows and columns 24
Partitioning via Kmeans... 25

Usage 27
Join... 27
Choosing distance metrics 28
Joining rows 29
Joining columns 29
Joining rows and columns 29
Labeling cases 30
Choosing linkage (amalgamation) methods 30
Saving files 31
Cluster displays 32
Kmeans... 34
Number of iterations 34
Number of groups 35
Labeling cases 35
Saving cluster membership 35

Computation 36
Memory limitations 36
Missing data 36
Algorithms 36

Examples 38
2.1 Joining with Euclidean distances 38
2.2 Joining with correlation metric 38
2.3 Joining a matrix 40
2.4 K-means 42
2.5 Distance matrix input 44

© 1989, SYSTAT, Inc.

Cluster 2

Overview

Cluster provides a variety of cluster analysis methods on rectangular or symmetric data matrices. Cluster analysis is a multivariate procedure for detecting natural groupings in data. It resembles discriminant analysis in one respect—the researcher seeks to classify a set of objects into subgroups although neither the number nor members of the subgroups are known.

Cluster provides two broad classes of clustering: **Join...** and **Kmeans...** The first class, **Join...**, comprises hierarchical, tree, or linkage methods. The second class, **Kmeans...**, comprises the K-means splitting method, which is not necessarily hierarchical.

Cluster/Join... operates on either rectangular SYSTAT files or files containing a symmetric matrix, such as those produced by **Corr**. First, you open a SYSTAT file containing a Rectangular, SSCP, Covariance, Correlation, Similarity, or Dissimilarity matrix. You can then select **Cluster/Join...** The other general clustering method, offered by **Cluster/Kmeans...**, works only on a rectangular matrix. **Data/By groups...** can be used to do separate analyses on subgroups of data, but only on rectangular matrices.

© 1989, SYSTAT, Inc.

Introduction

Cluster analysis is a multivariate procedure for detecting groupings in data. The objects in these groups may be:

1) **Cases** (observations or rows of a rectangular data file). For example, if health indicators (numbers of doctors, nurses, hospital beds, life expectancy, etc.) are recorded for countries (cases), then developed nations may form a subgroup or cluster separate from underdeveloped countries.

2) **Variables** (characteristics or columns of the data). For example, if causes of death (cancer, cardiovascular, lung disease, diabetes, accidents, etc.) are recorded for each U.S. state (case), the results show that accidents are relatively independent of the illnesses.

3) **Cases and variables** (individual entries in the data matrix). For example, certain wines are associated with good years of production. Other wines have other years that are better. This is because climate and growth of different grape varieties interact.

Types of clustering

Clusters may be of two sorts: *overlapping* or *exclusive*. Overlapping clusters allow the same object to appear in more than one cluster. Exclusive clusters do not. All the methods implemented in SYSTAT are exclusive.

There are two approaches to producing exclusive clusters: *hierarchical*, and *partitioned*. Hierarchical clusters consist of clusters that completely contain other clusters that completely contain other clusters, and so on. Partitioned clusters contain no other clusters. Hierarchical clusters can be represented by a tree. Partitioned cannot. In SYSTAT, hierarchical clusters are produced with the **Join...** procedure and partitioned clusters with the **Kmeans...** procedure.

© 1989, SYSTAT, Inc.

The cluster literature is diverse and contains many descriptive synonyms: hierarchical clustering (McQuitty, 1960; Johnson, 1967); single linkage clustering (Sokal & Sneath, 1963), and joining (Hartigan, 1975). Output from hierarchical methods can be represented as a tree (Hartigan, 1975) or dendrogram (Sokal & Sneath, 1963). (The linkage of each object or group of objects is shown as a joining of branches in a tree. The "root" of the tree is the linkage of all clusters into one set, and the ends of the branches lead to each separate object.)

Correlations and distances

To produce clusters, we must be able to compute some measure of dissimilarity between objects. Similar objects should appear in the same cluster and dissimilar objects in different clusters. All the methods available in **Corr** for producing matrices of association can be used in **Cluster**, but each has different implications for the clusters produced. Incidentally, **Cluster** converts correlations to dissimilarities by negating them.

In general, the correlation measures (**Pearson...**, **Mu2...**, **Spearman...**, **Gamma...**, **Tau...**) are not influenced by differences in scales between objects. For example, correlations between states using health statistics will not in general be affected by some states having larger average numbers or variation in their numbers. Use correlations when you wish to measure the similarity in *patterns* across profiles regardless of overall magnitude.

On the other hand, the other measures like **Euclidean...** and **City...** (city block distance) are significantly affected by differences in scale. For health data, two states will be judged different if they have differing overall incidences even when they follow a common pattern. You should use the distance measures generally when variables are measured on common scales.

© 1989, SYSTAT, Inc.

Standardizing data

Before you compute a dissimilarity measure, you may need to standardize your data across the measured attributes. Standardizing, you may recall, puts measurements on a common scale. In general, standardizing makes overall level and variation comparable across measurements. Consider the following data:

OBJECT	X1	X2	X3	X4
A	10	2	11	900
B	11	3	15	895
C	13	4	12	760
D	14	1	13	874

If we are clustering the four cases (A through D), then variable X4 will determine almost entirely the dissimilarity between cases whether we use correlations or distances. If we are clustering the four variables, then one of the correlation measures will adjust for the larger mean and standard deviation on X4. Thus, we should probably standardize within columns if we are clustering rows and use a correlation measure if we cluster columns.

Here is another example:

OBJECT	X1	X2	X3	X4
A	410	311	613	514
B	1	3	2	4
C	10	11	12	10
D	12	13	13	11

In this example, case A will have a disproportionate influence if we are clustering columns. We should probably standardize within rows before clustering columns. This requires using both **Transpose...** and **Standardize...** from the **Data** menu. If we are clustering rows, on the other hand, we should use a correlation measure to adjust for the larger mean and standard deviation of case A.

These are not immutable laws. The suggestions are only to make you realize that scales can influence distance *and* correlation measures.

© 1989, SYSTAT, Inc.

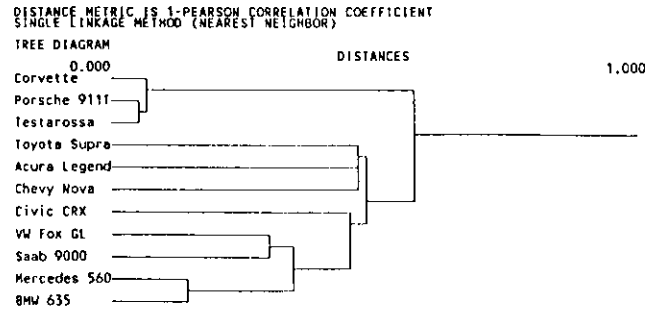
Hierarchical clustering

To understand hierarchical clustering, it's best to look at an example using **Join...** The following data are from the Overview in the **Corr** chapter. They are various attributes of selected performance cars.

ACCEL	BRAKE	SLALOM	MPG	SPEED	NAME$
5.0	245	61.3	17.0	153	Porsche 911T
5.3	242	61.9	12.0	181	Testarossa
5.8	243	62.6	19.0	154	Corvette
7.0	267	57.8	14.5	145	Mercedes 560
7.6	271	59.8	21.0	124	Saab 9000
7.9	259	61.7	19.0	130	Toyota Supra
8.5	263	59.9	17.5	131	BMW 635
8.7	287	64.2	35.0	115	Civic CRX
9.3	258	64.1	24.5	129	Acura Legend
10.8	287	60.8	25.0	100	VW Fox GL
13.0	253	62.3	27.0	95	Chevy Nova

Clustering rows

First, let's look at possible clusters of cars. Since the variables are on such different scales, we will standardize them before doing the clustering. This will give acceleration comparable influence to braking, for example. Then we select Pearson correlations as the basis for dissimilarity between cars. Here is the result.



If you look at the last correlation matrix in the Overview section of the **Corr** chapter, you will see how these clusters hang together. Cars within the same cluster (e.g. Corvette, Ferrari, Porsche) generally correlate highly.

© 1989, SYSTAT, Inc.

Clustering columns

We can cluster the performance attributes of the cars more easily. Here, we do not need to standardize within cars because all the values are comparable between cars. Again, to give each variable comparable influence, we will use Pearson correlations as the basis for the dissimilarities. Here is the result:

```
DISTANCE METRIC IS 1-PEARSON CORRELATION COEFFICIENT
SINGLE LINKAGE METHOD (NEAREST NEIGHBOR)
TREE DIAGRAM                    DISTANCES
     0.000                                        2.000
BRAKE
  MPG
ACCEL
SLALOM
SPEED
```

Clustering rows and columns

To cluster the rows and columns jointly, we should first standardize the variables to give each of them comparable influence on the clustering of cars. Once we have standardized, we can use Euclidean distances, because the scales are comparable. We used single linkage to produce the following result.

```
DISTANCE METRIC IS EUCLIDEAN DISTANCE
SINGLE LINKAGE METHOD (NEAREST NEIGHBOR)

                 S
              B AL S
              R CRP
              AKGLE
              KPEDE
              EGLND

Testarossa      :*
Porsche 911T    :*
Corvette        . :*
Acura Legend  : : *:
Toyota Supra  : : : :
BMW 635       : : : :
Saab 9000     : : : :
Mercedes 560  . : : .
VW Fox GL     *:*:
Chevy Nova    :**:
Civic CRX     **:*

LEGEND

          -1.952
                > BLANK
           -.583
                > .
            .785
                > *
           2.154

? = MISSING DATA OR ZERO
```


© 1989, SYSTAT, Inc.

This figure displays the standardized data matrix itself with rows and columns permuted to reveal clustering and each data value replaced by one of three symbols. Notice that the rows are ordered according to overall performance, with the fastest cars at the top. We chose only three symbols to display the values because this revealed the clusters most clearly.

Matrix clustering is especially useful for displaying large correlation matrices. You may want to cluster the correlation matrix this way and then use the ordering for a **SPLOM** in SYGRAPH to produce a scatterplot matrix that is organized by the multivariate structure.

Partitioning via Kmeans...

To produce partitioned clusters, you must decide in advance how many clusters you want. **Kmeans...** then searches for the best way to divide your objects into different sections so that they are separated as well as possible. **Kmeans...** begins by picking "seed" cases, one for each cluster, which are spread apart from the center of all the cases as much as possible. Then it assigns all cases to the nearest seed. Next, it attempts to reassign each case to a different cluster in order to reduce the within groups sum of squares. **Kmeans...** continues to reassign cases until the within groups sum of squares can no longer be reduced.

Kmeans... does not search through every possible partitioning of the data, so it is possible that some other solution might have smaller within-groups sums of squares. Nevertheless, it has performed relatively well on globular data separated in several dimensions in Monte Carlo studies of cluster algorithms.

Because it focuses on reducing within-groups sums of squares, **Kmeans...** is like a multivariate analysis of variance in which the groups are not known in advance. The output includes analysis of variance statistics, although you should be cautious in interpreting them. Remember, the program is looking for big *F*-ratios in the first place, so you should not be too impressed by large values.

Here is a three group **Kmeans...** analysis of our car data. The clusters are similar to those we found by joining. **Kmeans...** uses Euclidean distance instead of Pearson correlations, so there are minor differences due to scaling. To keep the influences of all variables comparable, we standardized the data before running **Kmeans...**

© 1989, SYSTAT, Inc.

Summary Statistics for 3 Clusters

Variable	Between SS	DF	Within SS	DF	F-Ratio	Prob
ACCEL	7.825	2	2.175	8	14.389	0.002
BRAKE	5.657	2	4.343	8	5.211	0.036
SLALOM	5.427	2	4.573	8	4.747	0.044
MPG	7.148	2	2.852	8	10.027	0.007
SPEED	7.677	2	2.323	8	13.220	0.003

Cluster Number: 1

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
Mercedes 560	0.60	ACCEL	-0.45	-0.14	0.17	0.23
Saab 9000	0.31	BRAKE	-0.15	0.23	0.61	0.28
Toyota Supra	0.49	SLALOM	-1.95	-0.89	0.11	0.73
BMW 635	0.16	MPG	-1.01	-0.47	-0.01	0.37
		SPEED	-0.34	0.00	0.50	0.31

Cluster Number: 2

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
Civic CRX	0.81	ACCEL	0.26	0.99	2.05	0.69
Acura Legend	0.67	BRAKE	-0.53	0.62	1.62	1.00
VW Fox GL	0.71	SLALOM	-0.37	0.72	1.43	0.74
Chevy Nova	0.76	MPG	0.53	1.05	2.15	0.65
		SPEED	-1.50	-0.91	-0.14	0.53

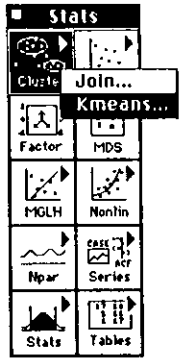
Cluster Number: 3

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
Porsche 911T	0.25	ACCEL	-1.29	-1.13	-0.95	0.14
Testarossa	0.43	BRAKE	-1.22	-1.14	-1.03	0.08
Corvette	0.31	SLALOM	-0.10	0.23	0.59	0.28
		MPG	-1.40	-0.78	-0.32	0.45
		SPEED	0.82	1.21	1.94	0.52

© 1989, SYSTAT, Inc.

Usage

To perform cluster analysis, you must specify the variables to be clustered, choose a type of clustering, and specify the details of the analysis method. **Cluster/Join...** provides hierarchical methods and **Cluster/Kmeans...** provides the K-means splitting method.



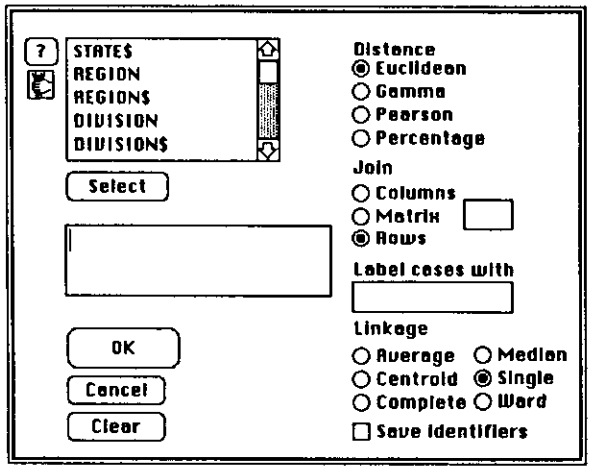
To use **Cluster**, press and hold the mouse button on **Cluster**. A submenu pops up offering **Join...** and **Kmeans...** Select the item you want and release the mouse button.

Join...

The **Join...** item produces hierarchical clusters that are displayed in a tree. You may cluster either rows or columns of your data or cluster a similarity or dissimilarity matrix directly.

- Select **Cluster/Join...** from the **Stats** menu
- Select variables to be analyzed
- Click OK

© 1989, SYSTAT, Inc.



Choosing distance metrics

If you use a rectangular data matrix, **Join...** computes normalized Euclidean distances between objects for its clustering metric. Normalized distances are root mean squared distances. Normalizing by the sample size allows comparison of clusterings across different size samples with or without missing data. Without missing data, normalized and regular Euclidean distances cluster identically.

[Euclidean]
This option uses normalized Euclidean distance (root mean squared distance). It is the default. Consider this metric for continuous or ratio scales.

[Gamma]
This option uses $1-g_{ij}$, where g_{ij} is the Goodman-Kruskal gamma correlation coefficient (see the **Corr** chapter). Consider this metric for rank order or ordinal scales.

[Pearson]
This option uses $1-p_{ij}$, where p_{ij} is the Pearson product-moment correlation between objects i and j. Consider this metric for continuous or interval scales.

© 1989, SYSTAT, Inc.

[Percentage]
This option produces a distance index that is the percentage of comparisons of values resulting in disagreements in two profiles. If two variables have the same values for all cases, for example, this index is zero when you join by columns. If they have different values on every case, this index is 100. Consider this metric for categorical or nominal scales.

All these distance indices are computed across columns if you join rows, and across rows if you join columns. Missing values are excluded from all computations. Profiles that are missing all their values are as far apart as possible.

See **Corr** for more distance options. With **Corr**, a variety of similarity and dissimilarity coefficients can be computed, including several that are popular for binary data. Matrices computed with **Corr** can be passed to **Cluster** for joining. **Cluster** automatically recognizes whether the matrix is similarity or dissimilarity. Just remember, if you want to cluster cases instead of variables, you must transpose the raw data matrix before computing a similarity or dissimilarity matrix in **Corr**.

Joining rows

Join... ordinarily clusters rows (cases) of data. Click the [Rows] button to reset to the default.

Joining columns

To cluster columns (variables), click the [Columns] button.

Joining rows and columns

To cluster rows and columns at the same time, which is really like doing a simultaneous cluster analysis for each, click [Matrix]. The display for matrix cluster analysis is different from the others. If you use the [Matrix] option with [Euclidean] distance (see Example 2.3), be sure that the variables are on comparable scales, since both rows and columns of the matrix are clustered. Joining a matrix containing inches of annual rainfall, annual growth of trees in feet, and proportion of deciduous trees per acre, for example, would split columns more by scales than by covariation.

© 1989, SYSTAT, Inc.

To display the clustered matrix, SYSTAT uses an adaptive routine to choose a small number of different symbols for displaying numerical intervals. This procedure, due to John Tukey, works well when the values in the matrix fall into easily separable groups (e.g. they are bimodal or trimodal). When the values in the matrix are uniformly distributed, you may want to control the number of symbols used to display the clustering. If you do this, the values in the matrix are split into as many equal width intervals as you choose. Just type the number of symbols (from 1 to 7) you want in the box next to the [Matrix] button. If you leave this box empty, SYSTAT chooses the number of symbols. Missing values will be displayed as question marks.

Labeling cases

Cases are denoted in tree and matrix displays with case numbers. If you have a character variable in your data file, you can label cases with its values. Just specify that variable in the [Label cases with] box.

Choosing linkage (amalgamation) methods

Various methods can be used to compute the distance of one object or cluster from another and to determine whether the two should be merged in a given step.

Single linkage

One ***amalgamation*** or linkage method takes the distance between two objects or clusters as the distance between the two closest members of those clusters. This method, the default, is often called ***single linkage clustering*** (Hartigan, 1975). It tends to produce long, stringy clusters. Choose this method by clicking the [Single] button. If you use a SYSTAT file that contains a Similarity or Dissimilarity matrix and choose [Single], you get what Johnson (1967) calls hierarchical clustering via the ***min*** method. If you use a rectangular matrix, this option produces a single linkage clustering.

Complete linkage

This linkage method uses the most distant pair of objects in two clusters to compute between-cluster distances. Complete linkage clustering tends to produce compact, globular clusters. If you use a Similarity or Dissimilarity matrix from a SYSTAT file and choose [Complete], you get Johnson's ***max*** method.

Centroid linkage

This linkage method uses the average value of all objects in a cluster (the cluster centroid) as the reference point for distances to other objects or clusters (Sokal and Michener, 1958).

© 1989, SYSTAT, Inc.

Average linkage

This linkage method averages all distances between pairs of objects in different clusters to decide how far apart they are (Sokal and Michener, 1958).

Median linkage

This method is similar to Average Linkage, except it uses medians (Gower, 1967).

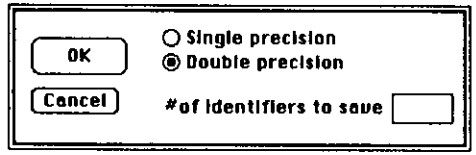
Ward's method

Ward's method (Ward, 1963) resembles Centroid Linkage, but adjusts for covariances.

For some data, the last four methods cannot produce a hierarchical tree with strictly increasing amalgamation distances. In these cases you may see stray branches that do not connect to others. If this happens, you should consider single or complete linkage. For more information on these problems, see Fisher and Van Ness (1971). These reviewers concluded that these and other problems made Centroid, Average, Median, and Ward (as well as K-means) "inadmissible" clustering procedures. In practice and in Monte Carlo simulations, however, they sometimes perform better than Single and Complete Linkage, which Fisher and Van Ness considered "admissible." Milligan (1980) tested all of the hierarchical joining methods in a large Monte Carlo simulation of clustering algorithms. You should consult his paper for further details.

Saving files

Click [Save identifiers] to save cluster membership identifiers in a file. A dialog box asks how many clusters to save identifiers for. You must be joining rows, not columns.



© 1989, SYSTAT, Inc.

For example, if you have 50 cases and 7 variables and you specify 6 clusters, SYSTAT joins the 50 cases 44 times to reduce the set of objects to 6 clusters and saves the integer identifiers for these 6 clusters into a SYSTAT file. If you don't know how many clusters are in your data, you should examine your tree display after a trial run to find relatively clear breaks between branches. Draw a vertical line at several positions along the tree and look for a cut that produces fairly distinct clumps of branches. Then choose the number of branches crossing your cut line for the desired number of clusters.

Only cluster identifiers are saved. You can merge these with your original variables in **Data**.

Cluster displays

There are several ways SYSTAT displays the output from **Join...** For joining rows or columns, SYSTAT prints a tree. For matrix joining, it prints a shaded matrix.

Trees

The tree is printed with a unique ordering in which every branch is lined up so that most similar objects are closest to each other. If a perfect *seriation* (one dimensional ordering) exists in the data, the tree produces it. The algorithm for ordering the tree is given in Gruvaeus and Wainer (1972). This ordering may differ from that of trees printed by other clustering programs if they do not use a seriation algorithm to determine how to order branches. The advantage of using seriation is most apparent for single linkage clusterings.

If you join rows, then the end branches of the tree are labeled with case numbers or labels (see [Label cases with] above). If you join columns, the end branches of the tree are labeled with variable names.

If you select [Show extended results] from **Data/Formats...**, SYSTAT displays the amalgamation distances or *cluster diameters* for each cluster at the right side of the tree. This helps to distinguish branches that appear to be tied at the same distance.

© 1989, SYSTAT, Inc.

Direct display of a matrix

The [Matrix] option of **Join...** produces a shaded display of the original data matrix in which rows and columns are permuted according to an algorithm in Gruvaeus and Wainer (1972). Different characters represent the magnitude of each number in the matrix (Ling, 1973). A legend showing the range of data values these characters represent appears below the display.

Cutpoints between these values and their associated characters are selected to heighten contrast in the display. The method for increasing contrast is derived from techniques used in computer pattern recognition, in which gray-scale histograms for visual displays are modified to heighten contrast and enhance pattern detection. To find these cutpoints, we sort the data and look for the largest gaps between adjacent values. Tukey's gapping method (Wainer and Schacht, 1978) is used to determine how many gaps (and associated characters) should be chosen to heighten contrast for a given set of data. This procedure, time consuming for large matrices, is described in detail in Wilkinson (1978).

If you have a course to grade and are looking for a way to find rational cutpoints in the grade distribution, you might want to use the [Matrix] display to choose the cutpoints. **Join...** the $n \times 1$ matrix of numeric grades (n students by 1 grade) and let SYSTAT choose the cutpoints. Only cutpoints asymptotically significant at the .05 level are chosen. If no cutpoints are chosen in the display, give everyone an A, flunk them all, or hand out numeric grades (unless you teach at Brown University or Hampshire College).

To control the number of symbols used in the [Matrix] display, specify a number (1–7) in the [Matrix] box. You may choose between 1 and 7 symbols for displaying data. If you use it, equal cutpoints between data values are chosen for the matrix display, rather than the contrast-optimized ones described above.

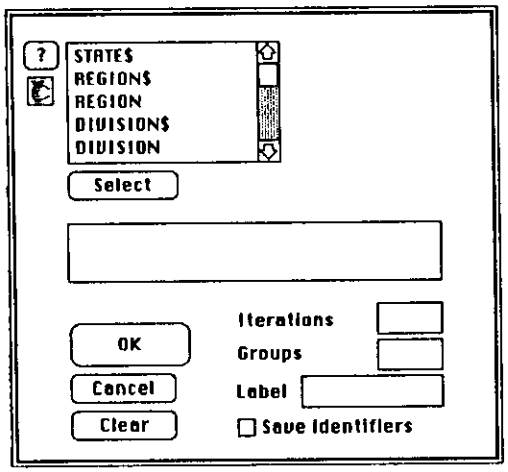
Missing data are displayed in the matrix as question marks. If the missing values are scattered randomly throughout the permuted data matrix, you can estimate missing values from adjacent values. However, if the missing values (question marks) cluster systematically, be cautious about estimating them with any procedure (including ones that exclude missing values). See "Missing data" in the "Computation" section below for more information on how to use the [Matrix] option of **Join...** to identify patterns of missing data.

© 1989, SYSTAT, Inc.

Kmeans...

The **Kmeans...** item produces a K-means clustering. You may need to transpose the data file before performing K-means clustering if you wish to cluster variables. K-means clustering splits a set of objects into a selected number of groups by maximizing between- relative to within-cluster variation. In rough terms it is like doing a one-way analysis of variance where the groups are unknown and the largest *F*-value is sought by reassigning members to each group. **Kmeans...** is an iterative procedure that assigns cases to a specified number of nonoverlapping clusters. It iterates through the data until it successfully clusters your cases. **Kmeans...** can only be used with a rectangular data matrix.

- Select **Cluster/Kmeans...** from the **Stats** menu
- Select variables to be analyzed
- Click OK



Number of iterations

You can specify the number of iterations **Kmeans...** uses to reallocate cases. The default value is 50. Rarely should you need more.

© 1989, SYSTAT, Inc.

Number of groups

You may change the number of groups formed with the [Number of groups] option. You may specify any integer value up to the number of cases in the file. Since a K-means clustering for one number of groups may not correspond to one for a different number, try several runs with different numbers. It is best to have a general idea of how many groups to expect in your data because adding clusters by increasing [Number of groups] eventually reveals more noise than information about your data.

Labeling cases

Cases are denoted in tree and matrix displays with case numbers. If you have a character variable in your data file, you can label cases with its values. Just specify that variable in the [Label cases with] box.

Saving cluster membership

Click [Save identifiers] in **Cluster/Kmeans...** to save a file of identifiers. You can merge the membership numbers with the original data in **Data**. See Saving files in the Join section directly above for a picture of the dialog box you get when clicking [Save identifiers].

© 1989, SYSTAT, Inc.

Computation

Memory limitations

Unlike most other SYSTAT routines, **Cluster** stores all the data in memory. Thus, if you have many cases (say, more than 200) you may get an "Out of memory" message when you attempt to join by columns even if you only have a few variables. The same is true when you have many variables and more than a few cases. To remedy this problem, compute a distance matrix with **Corr** and then use the matrix file for **Cluster**. All the distance measures available in the **Cluster** dialog box are also available with **Corr**.

In addition, you cannot **Join...** more cases or variables than the maximum number of variables allowed in a SYSTAT file (usually 256). If you have a file larger than this, split the file and join each half. This way you can check whether the clustering replicates.

Kmeans... can handle more cases than **Join...** It gives an "Out of memory" message when it chokes. If **Kmeans...** chokes, arrange for a decent burial.

Missing data

In **Cluster**, all distances are computed with pairwise deletion of missing values. See the **Corr** chapter for a description of this method. Distance matrices computed in other modules handle missing data according to the methods described in the relevant chapters.

Since missing data are excluded from distance calculations by pairwise deletion, they do not directly influence clustering when you use the [Matrix] option of the **Join...** command. To use the [Matrix] display to analyze patterns of missing data, create a new file in which missing values are recoded to 1 and all other values to 0. Then **Join...** with the [Matrix] option to see whether missing values cluster together in a systematic pattern.

Algorithms

All computations are done in single precision. The options for **Join...** are computed in a single subroutine that follows the standard hierarchical amalgamation method described in Hartigan (1975). The algorithm in Gruvaeus and Wainer (1972) is used to order the tree.

© 1989, SYSTAT, Inc.

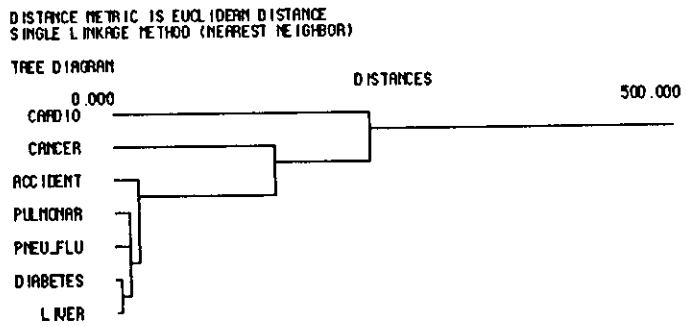
Kmeans... follows the algorithm described in Hartigan (1975). Modifications from Hartigan and Wong (1979) improve speed. There is an important difference between SYSTAT's **Kmeans...** algorithm and that of Hartigan (or implementations of Hartigan's in BMDP, SAS, and SPSS). In SYSTAT, seeds for new clusters are chosen by finding the case farthest from the centroid of all cases in Euclidean distance. In Hartigan's algorithm, seeds for new clusters are chosen by splitting on the variable with largest variance.

© 1989, SYSTAT, Inc.

2.1 Joining with Euclidean distances

Here is a cluster analysis of death-by-cause variables in the MEDICAL data. The cases are U.S. states. We use the defaults [Single] linkage and [Euclidean] distances.

- Select **Cluster/Join...** from the **Stats** menu
- Select the variables to be analyzed: ACCIDENT, CARDIO, CANCER, PULMONAR, PNEU_FLU, DIABETES, LIVER
- Choose [Columns]
- Click OK



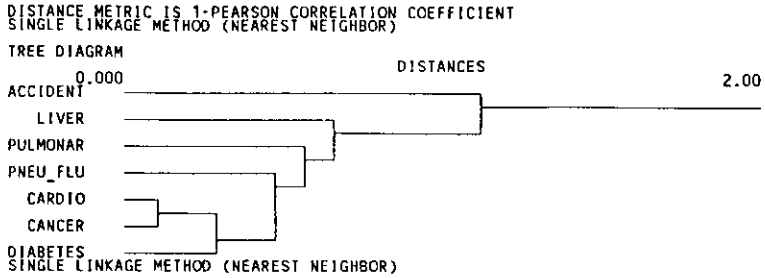
Notice that CARDIO and CANCER are the last variables to join the cluster tree, as if their incidence were unrelated to that of the other medical classifications.

2.2 Joining with correlation metric

Since we used Euclidean distances, the structure of the previous tree was influenced by the scales of the variables. Let's try to remove grouping due to scales with Pearson correlations, which standardize each variable and thus force the weights to be equal.

- Select **Cluster/Join...** from the **Stats** menu
- Select variables to be analyzed: ACCIDENT, CANCER, CARDIO, PNEU_FLU, PULMONAR, DIABETES, LIVER
- Choose [Columns]
- Choose [Pearson]
- Click OK

© 1989, SYSTAT, Inc.



Notice in this solution that CANCER and CARDIO, which in the first solution did not cluster with the other variables, are now the first pair of variables to cluster and associate with other illnesses. Accidents are now independent of the illness variables.

© 1989, SYSTAT, Inc.

2.3 Joining a matrix

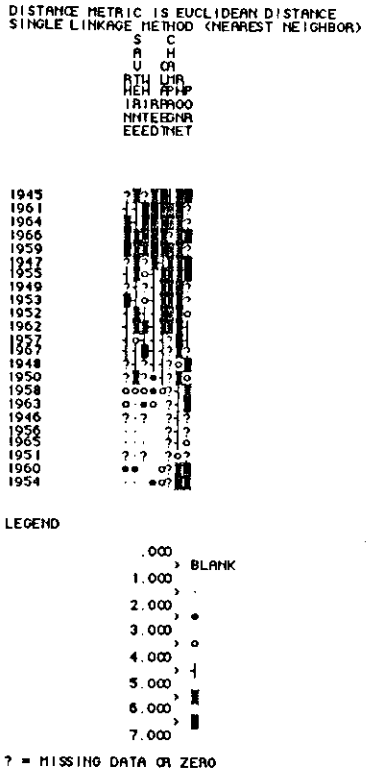
Here we join an entire data matrix (rows and columns). These data (WINES) were prepared by the International Wine and Food Society, cited in Simon (1969). They show ratings of wines by year (0=poor; 7=outstanding). RED and WHITE are red and white Burgundies, respectively.

CLARET	RED	WHITE	SAUTERNE	RHONE	RHINE	CHAMPAGN	PORT	YEAR$
7	6	.	6	7	.	5	7	1945
1	1	.	2	5	.	.	.	1946
5	6	.	6	6	.	6	7	1947
5	5	.	5	4	.	.	7	1948
6	5	.	5	7	.	6	.	1949
5	3	.	6	6	.	.	4	1950
0	1	.	2	4	.	.	.	1951
6	5	5	6	6	5	7	4	1952
6	5	4	5	6	7	6	.	1953
4	3	1	2	7	2	.	6	1954
6	5	4	6	6	5	6	7	1955
0	1	1	2	5	2	.	.	1956
5	5	5	4	7	5	.	5	1957
4	3	4	4	5	4	.	6	1958
6	7	6	6	6	7	7	.	1959
4	1	1	3	6	3	.	7	1960
7	7	7	5	7	5	7	.	1961
6	5	6	6	6	5	6	5	1962
1	4	3	2	5	4	.	7	1963
6	7	7	5	6	6	7	.	1964
0	1	2	2	5	2	.	4	1965
6	7	7	6	6	6	.	7	1966
5	5	7	5	6	5	.	.	1967

We will cluster this matrix directly. To make the display more readable, we've chosen the generic symbol set with 7 symbols.

- Select **Cluster/Join...** from the **Stats** menu
- Choose |Matrix| and specify 7 symbols
- Specify character variable for the labels: YEAR$
- Click OK

© 1989, SYSTAT, Inc.



This clustering reveals the distinctive data structure of a Guttman simplex (Guttman, 1944; 1954). The seriation algorithm for ordering the row and column trees, discussed above, produces a unidimensional scale of wine years. Notice that the whole matrix looks dark in the upper triangular portion and light in the lower. The best vintage years have been permuted to the top of the data matrix; wines with the best ratings over the years have been permuted to the right.

© 1989, SYSTAT, Inc.

The Guttman simplex is not a necessary consequence of the two-way clustering and permutation algorithm. Data frequently have an irregular or stepwise structure. The multitrait-multimethod matrix of Campbell and Fiske (1959) is a good example. If you permute a symmetric matrix of correlations from **Corr**, you may encounter this structure.

2.4 K-means

Let's do a K-means clustering of American states based on the medical variables in USDATA. We choose 4 clusters for this display. To keep the influence of the variables comparable, we first standardized the variables in **Data**.

- Select **Cluster/Kmeans...** from the **Stats** menu
- Select the variables to be analyzed: ACCIDENT, CANCER, CARDIO, LIVER, PNEU_FLU, PULMONAR, DIABETES
- Specify the [Number of groups]: 4
- Click OK

Summary Statistics for 4 Clusters

Variable	Between SS	DF	Within SS	DF	F-Ratio	Prob
ACCIDENT	25.980	3	23.020	46	17.306	0.000
CARDIO	29.796	3	19.204	46	23.791	0.000
CANCER	31.683	3	17.317	46	28.054	0.000
PULMONAR	24.316	3	24.684	46	15.104	0.000
PNEU_FLU	17.908	3	31.092	46	8.831	0.000
DIABETES	19.863	3	29.137	46	10.453	0.000
LIVER	22.612	3	26.388	46	13.139	0.000

Cluster Number: 1

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
NH	0.57	ACCIDENT	-0.96	0.02	1.66	0.69
VT	0.55	CARDIO	-0.80	0.42	1.24	0.57
IN	0.63	CANCER	-0.39	0.23	0.99	0.34
WI	0.70	PULMONAR	-0.71	0.67	1.75	0.63
MN	0.66	PNEU_FLU	-0.59	0.65	2.74	0.80
IA	0.62	DIABETES	-0.99	0.01	0.97	0.54
MO	0.40	LIVER	-1.22	-0.39	0.93	0.60
SD	0.97					
NB	0.56					
KS	0.38					
WV	0.70					
KY	0.23					
TN	0.48					
AR	0.40					
OK	0.65					
MT	0.68					
AZ	0.90					
WA	0.54					
OR	0.53					

© 1989, SYSTAT, Inc.

Cluster Number: 2

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
AK	0.00	ACCIDENT	3.78	3.78	3.78	0.00
		CARDIO	-3.37	-3.37	-3.37	0.00
		CANCER	-3.10	-3.10	-3.10	0.00
		PULMONAR	-3.25	-3.25	-3.25	0.00
		PNEU_FLU	-2.09	-2.09	-2.09	0.00
		DIABETES	-3.12	-3.12	-3.12	0.00
		LIVER	0.16	0.16	0.16	0.00

Cluster Number: 3

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
ND	0.89	ACCIDENT	-1.08	0.39	2.12	0.84
MD	0.79	CARDIO	-2.16	-0.76	0.28	0.77
VA	0.54	CANCER	-2.56	-0.77	0.50	0.80
NC	0.38	PULMONAR	-1.87	-0.56	1.04	0.84
SC	0.36	PNEU_FLU	-1.77	-0.57	1.17	0.74
GA	0.29	DIABETES	-1.54	-0.40	0.81	0.68
AL	0.63	LIVER	-1.58	-0.44	1.77	0.78
MS	0.68					
LA	0.72					
TX	0.28					
ID	0.59					
WY	1.01					
CO	0.82					
NM	1.02					
UT	1.08					
NV	1.20					
HI	1.07					

Cluster Number: 4

Members		Statistics				
Case	Distance	Variable	Minimum	Mean	Maximum	St.Dev.
ME	0.55	ACCIDENT	-1.43	-0.84	0.15	0.40
MA	0.85	CARDIO	-0.48	0.63	1.31	0.49
RI	1.00	CANCER	-0.16	0.92	1.98	0.57
CT	0.46	PULMONAR	-0.64	-0.00	1.34	0.62
NY	0.79	PNEU_FLU	-1.22	-0.04	2.09	0.86
NJ	0.29	DIABETES	-1.13	0.75	3.10	1.10
PA	0.48	LIVER	-0.13	1.13	2.61	0.84
OH	0.58					
IL	0.42					
MI	0.51					
DE	1.02					
FL	0.96					
CA	1.04					

© 1989, SYSTAT, Inc.

The summary *F*-Ratios at the beginning of the output indicate that CARDIO and CANCER are better discriminators between states than other illnesses such as pneumonia and influenza. The probabilities associated with these *F* values have to be taken with a grain of salt because the clustering optimized them for the number of groups. You might want to take the three variables with the highest *F*-Ratios (CARDIO, CANCER, and ACCIDENT) and rotate the data in 3-D to look for the clusters.

Notice that Alaska, which has extremely low illness rates, is an outlier that forms its own cluster. This is not necessarily undesirable, because it allows clustering to proceed on the remaining cases without being influenced by Alaska. See the **Factor** chapter for a plot of the first two rotated components of this dataset. The K-means clustering fits this plot quite well.

2.5 Distance matrix input

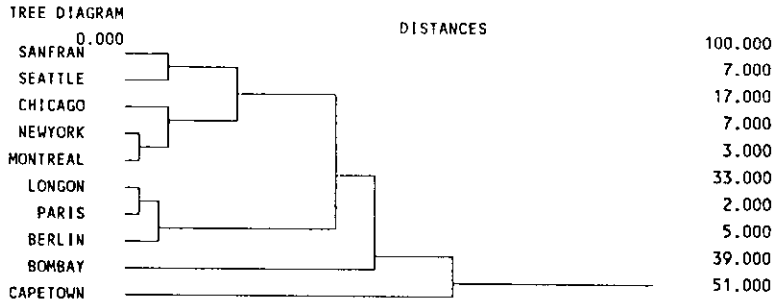
Here is an example of clustering a matrix of distances that were entered directly into the SYSTAT editor. The data are airline distances in hundreds of miles between ten global cities. The rows are the same as the columns. When you enter these numbers in the editor, you must specify their type as dissimilarity. The data are adapted from Hartigan (1975). We chose [Show extended results] in order to see the amalgamation distances.

BERLIN	BOMBAY	CAPETOWN	CHICAGO	LONDON	MONTREAL	NEWYORK	PARIS	SANFRAN	SEATTLE
0									
39	0								
60	51	0							
44	81	85	0						
6	45	60	40	0					
37	75	79	8	33	0				
40	78	78	7	35	3	0			
5	44	58	41	2	34	36	0		
57	84	103	19	54	25	26	56	0	
51	77	102	17	48	23	24	50	7	0

© 1989, SYSTAT, Inc.

- Select **Data/Formats...**
- Choose [Show extended results]
- Click OK

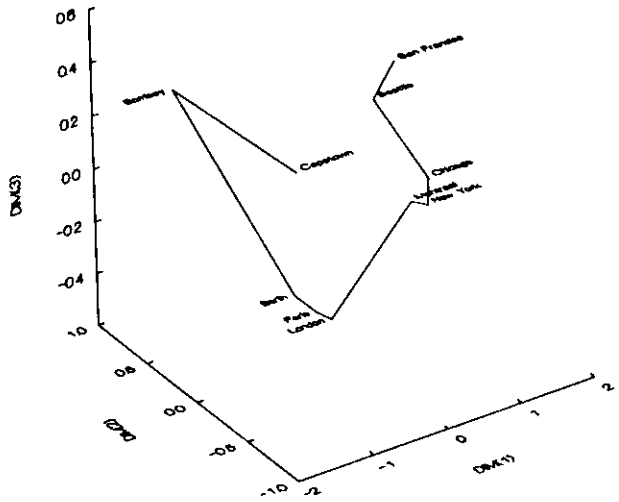
- Select **Cluster/Join...** from the **Stats** menu
- Specify the variables to be analyzed: BERLIN, BOMBAY, CAPETOWN, CHICAGO, LONDON, MONTREAL, NEWYORK, PARIS, SANFRAN, SEATTLE
- Click OK



The tree is printed in seriation order. Imagine a trip around the globe to these cities. SYSTAT has identified a shortest path between cities. The itinerary begins at San Francisco, leads to Seattle, then to Chicago, to New York, and so on, until ending in Capetown.

© 1989, SYSTAT, Inc.

We have a little problem with the airline distances, however. They are on the surface of the globe. Spatially, we should be working with the distances *through* the globe (chords). Here is a multidimensional scaling of the same data in three dimensions. We have used the [Span] option in *Graph* to delineate the minimum spanning tree in the three dimensional solution. Now you can see the path that **Cluster** chose. If we had the real three dimensional coordinates of the cities, our path would be a better representation.



We used this example both to illustrate direct distance input and to give you an idea of the kind of information contained in the order of the SYSTAT cluster tree. For distance data, the seriation reveals shortest paths. And for typical sample data, the seriation is more likely to replicate in new samples so that you can recognize cluster structure.

© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

3 Corr

Introduction 50
The scatterplot matrix (SPLOM) 50
The Pearson correlation coefficient 51
Other measures of association 52
Measures for rank order data 52
Measures for binary data 53
Transposed data 55
Pairwise and listwise deletion of missing data 56
Usage 58
City... 58
Covariance... 59
Dichotomous... 59
Euclidean... 60
Gamma... 60
Mu2...
Pearson... 61
Probabilities for Pearson correlation coefficients 62
SSCP... 62
Spearman... 62
Tau... 63
Saving files 63
Computation 64
Missing data 64
Algorithms 64
Examples 65
3.1 Pearson correlation 65
3.2 Probabilities 66
3.3 Bonferroni-adjusted probabilities 67

© 1989, SYSTAT, Inc.

Corr 3

Overview

Corr computes symmetric matrices of correlation and similarity on a data set. It prints the resulting matrix and, if you wish, saves it into a SYSTAT file for further analysis, such as multidimensional scaling, cluster, or factor analysis.

© 1989, SYSTAT, Inc.

The **Corr** procedures compute many different measures of the strength of association between variables. The most popular measure, and the one after which **Corr** is named, is the Pearson correlation. The *Pearson correlation coefficient* varies between –1 and +1. A Pearson correlation of 0 indicates that neither of two variables can be predicted from the other by using a linear equation. A Pearson correlation of 1 indicates that one variable can be predicted perfectly by a positive linear function of the other, and vice versa. And a value of –1 indicates the same, except that the function has a negative sign for the slope of the line. More generally, **Corr** computes a variety of similarity and distance measures.

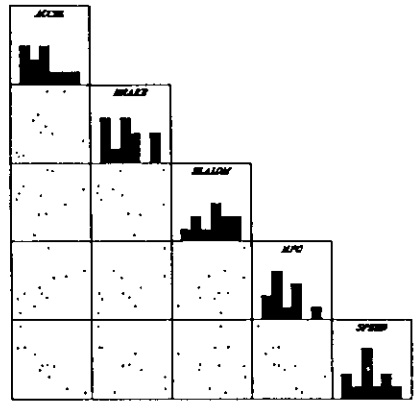
Let's look at an example. The following data are taken from various issues of *Car and Driver* and *Road & Track* magazine. They are the car enthusiasts' equivalent of *Consumer Reports* performance ratings. The cars rated include some of the most expensive and exotic cars in the world (e.g. Ferrari Testarossa) as well as some of the least expensive but sporty cars (e.g. Honda Civic CRX). The attributes measured are 0–60 m.p.h. acceleration, braking distance in feet from 60 to 0 m.p.h., slalom times (speed over a twisty course), miles-per-gallon, and top speed in miles-per-hour.

ACCEL	BRAKE	SLALOM	MPG	SPEED	NAME$
5.0	245	61.3	17.0	153	Porsche 911T
5.3	242	61.9	12.0	181	Testarossa
5.8	243	62.6	19.0	154	Corvette
7.0	267	57.8	14.5	145	Mercedes 560
7.6	271	59.8	21.0	124	Saab 9000
7.9	259	61.7	19.0	130	Toyota Supra
8.5	263	59.9	17.5	131	BMW 635
8.7	287	64.2	35.0	115	Civic CRX
9.3	258	64.1	24.5	129	Acura Legend
10.8	287	60.8	25.0	100	VW Fox GL
13.0	253	62.3	27.0	95	Chevy Nova

The scatterplot matrix (SPLOM)

A convenient summary that shows the relationship between the performance variables is to arrange them in a *matrix*. A matrix is a rectangular array. In the cells of the matrix we may put any sort of numbers, but we will focus on measures of association. Before doing that, however, let's examine a graphical matrix, the *scatterplot matrix* (SPLOM).

© 1989, SYSTAT, Inc.



This matrix shows the histograms of each variable on the diagonal and the *scatterplots* (*x*-*y* plots) of each variable against the others. For example, the scatterplot of acceleration vs. braking is at the top of the matrix. Since the matrix is *symmetric*, only the bottom half is shown. In other words, the plot of acceleration vs. braking is the same as the transposed scatterplot of braking vs. acceleration.

The Pearson correlation coefficient

Now, assume we want a single number that summarizes how well we could predict acceleration from braking using a straight line. The "Introduction" of the **MGLH** chapter discusses how we calculate such a line, but it is enough here to know that we are interested in drawing a line through the area covered by the points in the scatterplot such that, on average, the acceleration of a car could be predicted rather well by the value on the line corresponding to its braking. The closer the points cluster around this line, the better would be the prediction.

In addition, we want this number to represent simultaneously how well we can predict braking from acceleration using a similar line. This symmetry we seek is fundamental to all the measures available in **Corr**. It means that, whatever the scales on which we measure our variables, the coefficient of association we compute will be the same for either prediction. If this symmetry makes no sense for a certain dataset, then you should probably not be using **Corr**.

© 1989, SYSTAT, Inc.

Here is the Pearson correlation matrix corresponding to this SPLOM:

Pearson Correlation Matrix

	ACCEL	BRAKE	SLALOM	MPG	SPEED
ACCEL	1.000				
BRAKE	0.466	1.000			
SLALOM	0.176	-0.097	1.000		
MPG	0.651	0.622	0.597	1.000	
SPEED	-0.908	-0.665	-0.115	-0.768	1.000

Number of Observations: 11

Try superimposing the correlation matrix on the SPLOM in your mind. The Pearson correlation for acceleration vs. braking is .466. This correlation is positive and moderate in size. On the other hand, the correlation between acceleration and speed is negative and quite large (–.908). You can see in the lower left corner of the SPLOM that the points cluster around a downward sloping line. In fact, all the correlations of speed with the other variables are negative, which makes sense since greater speed implies greater performance. The same is true for slalom performance, but this is clouded by the fact that some small but slower cars like the Honda Civic CRX are extremely agile.

The Pearson correlation measures *linear* predictability. Do not assume that a Pearson correlation near zero implies no relationship between variables. Many nonlinear associations (U- and S-shaped curves, for example) can have Pearson correlations of zero.

Other measures of association

Corr offers a variety of other association measures. There is not room here to discuss all of them, but let's review some briefly.

Measures for rank order data

Several measures are available for rank order data. If your data are simply ranks of attributes, or you wish to see how well variables are associated when you pay attention only to rank orderings, then you should consider these measures. They are: Goodman-Kruskal gamma, Guttman mu2, Spearman rho, and Kendall tau. Each measures an aspect of rank order association. The one closest to Pearson is the Spearman. The Spearman rho is simply a Pearson correlation computed on the same data after converting them to ranks.

© 1989, SYSTAT, Inc.

Here is the same matrix computed for Spearman rho:

Matrix of Spearman Correlation Coefficients

	ACCEL	BRAKE	SLALOM	MPG	SPEED
ACCEL	1.000				
BRAKE	0.501	1.000			
SLALOM	0.245	-0.305	1.000		
MPG	0.815	0.502	0.487	1.000	
SPEED	0.891	-0.651	-0.109	-0.884	1.000

Number of observations: 11

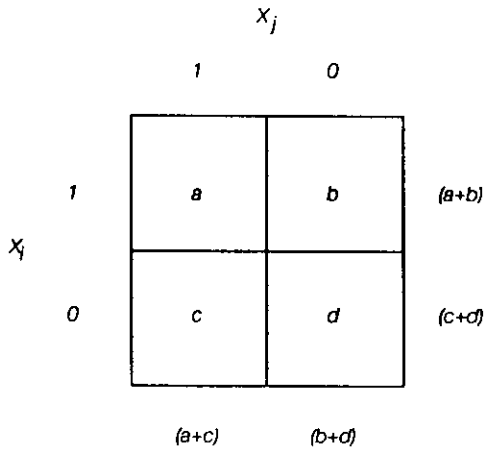
It is often useful to compute both a Spearman and Pearson matrix on the same data. The absolute difference between the two can reveal unusual features. For example, the greatest difference for our data is on the slalom-braking correlation. This is because the Honda Civic CRX is so fast through the slalom, despite its inferior brakes, that it attenuates the Pearson correlation between slalom and braking. The Spearman correlation reduces its influence.

Measures for binary data

Corr also offers association measures for binary data. These coefficients relate variables whose values may represent presence or absence of an attribute or simply two values. The similarity coefficients S2–S6 work only for dichotomous data scored as 0 or 1. They are documented in Gower (1985). The names of these coefficients follow Gower's table on page 309. These coefficients were chosen for SYSTAT because they are *metric* (see the **MDS** introduction for a definition) and produce symmetric positive semi-definite (Gramian) matrices, provided you do not use the pairwise deletion option. This makes them suitable for multidimensional scaling and factoring as well as clustering.

© 1989, SYSTAT, Inc.

The following table shows how the similarity coefficients are computed:



S2 = $\frac{a}{a+b+c+d}$ proportion of pairs with both values present

S3 = $\frac{a}{a+b+c}$ proportion of pairs with both values present *given* at least one occurs

S4 = $\frac{a+d}{a+b+c+d}$ proportion of pairs where the values of both variables agree

S5 = $\frac{a}{a+2(b+c)}$ S3 standardized by all possible patterns of agreement and disagreement

S6 = $\frac{a+d}{a+2(b+c)+d}$ S4 standardized by all possible patterns of agreement and disagreement

When the absence of an attribute in both variables is deemed to convey no information, then *d* should not be included in the coefficient (see S3 and S5)

© 1989, SYSTAT, Inc.

Transposed data

You can use **Corr** to compute measures of association on the rows or columns of your data. Simply transpose the data using **Data/Transpose...**and then use **Corr**. This makes sense when you wish to assess similarity between rows. We might be interested in identifying similar cars from our performance measures, for example.

When you compute association measures across rows, however, be sure that the variables are on comparable scales. Otherwise, a single variable will influence most of the association. With the cars data, braking and speed are so large that they would almost uniquely determine the similarity between cars. Consequently, we standardized the data before transposing them. That way, the correlations measure the similarities comparably across attributes.

Here is the Pearson correlation matrix for our cars:

```
Pearson Correlation Matrix

                PORSCHE     FERRARI     CORVETTE    MERCEDES        SAAB

PORSCHE           1.000
FERRARI           0.940       1.000
CORVETTE          0.939       0.868       1.000
MERCEDES          0.093       0.212      -0.240       1.000
SAAB             -0.506      -0.523      -0.760       0.664       1.000
TOYOTA            0.238       0.429       0.402      -0.379      -0.680
BMW              -0.319      -0.095      -0.557       0.854       0.634
HONDA            -0.504      -0.730      -0.393      -0.519       0.765
ACURA            -0.046      -0.102       0.298      -0.978       0.770
VW               -0.962      -0.928      -0.980       0.079       0.704
CHEVY            -0.731      -0.698      -0.491      -0.532      -0.131

                TOYOTA          BMW       HONDA       ACURA          VW

TOYOTA            1.000
BMW              -0.247       1.000
HONDA            -0.298      -0.500       1.000
ACURA             0.533      -0.788       0.349       1.000
VW               -0.353       0.391       0.552      -0.156       1.000
CHEVY            -0.034      -0.064       0.320       0.536       0.525

                CHEVY

CHEVY             1.000

Number of observations:    5
```

You may want to check the introduction of the **Cluster** chapter to see how similar cars cluster in this matrix.

© 1989, SYSTAT, Inc.

Pairwise and listwise deletion of missing data

If you have missing data, Corr can handle them in two ways: listwise or pairwise deletion. Listwise deletion is the default. If there are missing data and pairwise deletion is used, SYSTAT displays a table of frequencies between all possible pairs of variables after the correlation matrix.

Listwise deletion removes from computations any observation with a value missing on any variable included in the correlation matrix requested.

Pairwise deletion is listwise deletion done separately for every pair of selected variables. In other words, counts, sums of squares, and sums of cross products are computed separately for every pair of variables in the file. With pairwise deletion, you get the same correlation (covariance, etc.) for two variables containing missing data if you select them alone or with other variables containing missing data. With listwise deletion, correlations under these two circumstances may differ, depending on the pattern of missing data among the other variables in the file.

Pairwise deletion takes considerably more computer time because the sums of cross products for each pair must be saved in a temporary disk file. If you use the pairwise deletion to compute an SSCP matrix, the sums of squares and cross products are weighted by N/n, where N is the number of cases in the whole file and n is the number of cases with nonmissing values in a given pair.

Perhaps because it is so convenient, pairwise deletion is a popular method for computing correlations on matrices with missing data. Many regression programs include it as a standard method for computing regression estimates from a covariance or correlation matrix.

Ironically, pairwise deletion is one of the worst ways to handle missing values. If as few as 20% of the values in a data matrix are missing, it is not difficult to find two correlations that were computed using substantially different subsets of the cases. In such cases, it is common to encounter error messages that the matrix is singular in regression programs (see MGLH) and to get eigenvalues less than zero (see Factor).

© 1989, SYSTAT, Inc.

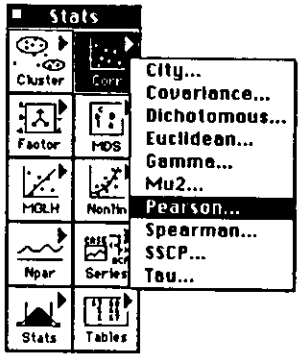
If you have missing values, you will do better by taking the time to impute (replace) them in the raw data by examining similar cases or variables with nonmissing values. The [Matrix] option in **Cluster/ Join...** offers one way to do this. Another way is to compute a regression equation to predict each variable from non-missing data on other variables and then to use that equation to predict missing values. For more information on the missing value problem, see Beale and Little (1975).

© 1989, SYSTAT, Inc.

Usage

The **Corr** item in the **Stats** menu provides various types of correlation and similarity matrices according to the subitem you choose:

City...	city block distances
Covariance...	covariances
Dichotomous...	five types of dichotomy coefficients
Euclidean...	normalized Euclidean distances
Gamma...	*gamma* coefficients
Mu2...	Guttman *mu2* coefficients
Pearson...	Pearson correlations
Spearman...	Spearman rank correlations
SSCP...	sums of squares and cross products
Tau...	Kendall *tau-b* coefficients



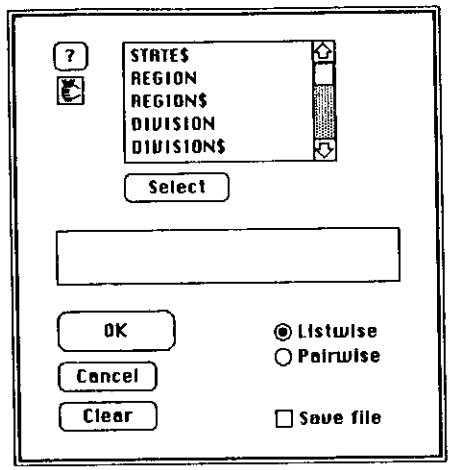
To use **Corr**, press and hold the mouse button on **Corr**. A submenu pops up offering the various types of correlation. Select the item you want and release the mouse button.

City...

This computes city block distances (sum of absolute discrepancies).

- Select **Corr/City...** from the **Stats** menu
- Select variables; all numeric variables are used by default
- Click OK

© 1989, SYSTAT, Inc.



City block distances are like Euclidean distances, except absolute differences are used instead of squared distances.

Covariance...

This computes sample variances and covariances (using N-1) in the denominator.

- Select **Corr/Covariance...** from the **Stats** menu
- Select variables; all numeric variables are used by default
- Click OK

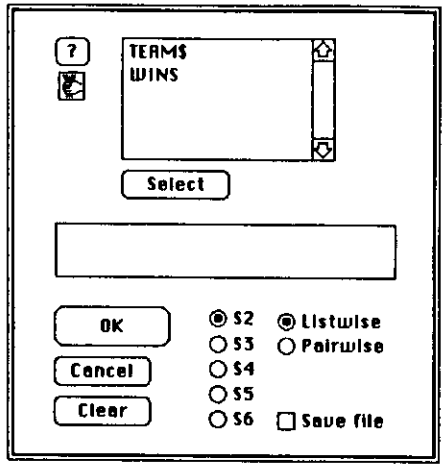
The dialog box is the same as that for **City...**, shown above.

Dichotomous...

This computes coefficients for binary data (see Introduction).

- Select **Corr/Dichotomous...** from the Stats menu
- Choose type of dichotomy coefficient: S2, S3, S4, S5, or S6
- Select variables; all numeric variables are used by default
- Click OK

© 1989, SYSTAT, Inc.



[S2] computes positive matching dichotomy coefficients, [S3] computes Jaccard's dichotomy coefficients, [S4] computes simple matching dichotomy coefficients, [S5] computes Anderberg's dichotomy coefficients, and [S6] computes Tanimoto's dichotomy coefficients.

Euclidean...

This computes root mean squared Euclidean distances.

- Select **Corr/Euclidean...** from the Stats menu
- Select variables; all numeric variables are used by default
- Click OK

The normalized Euclidean distances are root mean squared distances, which keeps them from getting out of hand when the sample size is large. The dialog box is the same as that for **City...**, shown above.

Gamma...

This computes Goodman-Kruskal *gamma* coefficients.

- Select **Corr/Gamma...** from the Stats menu
- Select variables; all numeric variables are used by default
- Click OK

See Goodman and Kruskal (1954) for the *gamma* coefficient. The dialog box is the same as that for **City...**, shown above.

© 1989, SYSTAT, Inc.

Mu2...

This computes Guttman's *mu2* monotonicity coefficients.

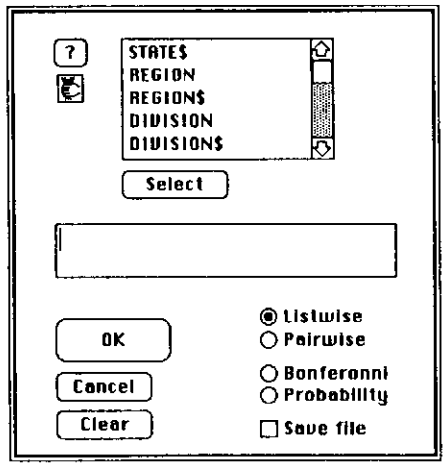
- Select **Corr/Mu2...** from the Stats menu
- Select variables; all numeric variables are used by default
- Click OK

See Shye (1978) for documentation of the *mu2* monotonicity coefficient. The dialog box is the same as that for **City...**, shown above.

Pearson...

This computes the usual Pearson product moment correlation coefficients.

- Select **Corr/Pearson...** from the **Stats** menu
- Select variables; all numeric variables are used by default
- Choose [Bonferonni] or [Probability] for probabilities associated with each coefficient
- Click OK



© 1989, SYSTAT, Inc.

Probabilities for Pearson correlation coefficients

Corr/Pearson... has two additional options, [Bonferroni] and [Probabilities]. [Probabilities] displays the probabilities associated with each correlation coefficient; [Bonferroni] displays the Bonferroni-adjusted probabilities. In addition, both produce a Bartlett chi-square test of whether the population correlation matrix is an identity, i.e. a global test of whether the set of correlations is insignificant:

$$\chi^2 = -\left\{N - 1 - \frac{(2p+5)}{6}\right\} \ln |\mathbf{R}|$$

where N is the total sample size (or the smallest sample size for any pair in the matrix if pairwise deletion is used), p is the number of variables, and $|\mathbf{R}|$ is the determinant of the correlation matrix. This test is sensitive to non-normality and the test statistic is only asymptotically distributed (for large samples) as chi-square. Nevertheless, it can serve as a rough guideline. See Example 3.2 for more information.

SSCP...

This computes sums of squares and cross products.

- Select **Corr/SSCP...** from the **Stats** menu
- Select variables; all numeric variables are used by default
- Click OK

The dialog box is the same as that for **City...**, shown above.

Spearman

This computes Spearman rank correlation coefficients.

- Select **Corr/Spearman...** from the **Stats** menu
- Select variables; all numeric variables are used by default
- Click OK

Note: another way to produce a Spearman rank order correlation matrix is first to use **Data/Rank...** on your data file and then use **Corr/Pearson...**

The dialog box is the same as that for **City...**, shown above.

© 1989, SYSTAT, Inc.

Tau...

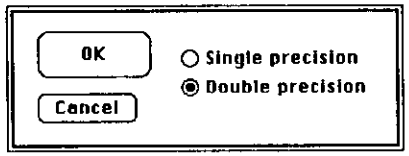
This computes Kendall *tau* coefficients.

- Select **Corr/Tau...** from the Stats menu
- Select variables; all numeric variables are used by default
- Click OK

The dialog box is the same as that for **City...**, shown above.

Saving files

Click [Save file] to save the resulting correlation matrix in a file. A dialog box asks whether to save in single or double precision.



The similarity coefficients are stored in a SYSTAT file with type [Similarity]. City and Euclidean distances are stored with type [Dissimilarity]. The correlation coefficients are stored with type [Correlation] and the remaining matrices are stored with their own types. (To find out what type a data file has, either open the file in the Data Editor and check in **Editor/Formats...** to see which button is on, or use **Show command window** before opening the file; when you open a file, the Command window shows its type and the names of the variables it includes.)

© 1989, SYSTAT, Inc.

Computation

Missing data

See the "Introduction" for information on the treatment of missing values. If there are missing data and the [Pairwise] option is used, SYSTAT prints a table of frequencies between all possible pairs of variables after printing the correlation matrix.

Algorithms

The computational algorithms use provisional means, sums of squares, and cross products (Spicer, 1972). All are implemented in double precision. The **City...**, **Covariance...**, **Dichotomous ..**, **Euclidean...**, **Pearson...**, and **SSCP...** routines can handle unlimited observations because they use available memory plus scratch disk files to contain the provisional calculations. The larger the requested matrix, the longer the calculations take, because less memory is available for buffering.

For the rank order coefficients (**Gamma...**, **Mu2...**, **Spearman...**, and **Tau...**), the raw data must fit in memory. If you have too many variables and observations, SYSTAT reports that it is out of memory. Even on small machines, however, fairly large matrices can be handled. but keep in mind that these rank order coefficients are time consuming. **Spearman...** requires sorting and ranking the data before doing the same work done by **Pearson...** The **Gamma...** and **Mu2...** items require computations between all possible pairs of observations. Thus, their computing time is combinatoric and becomes unreasonable after 100 observations or so.

© 1989, SYSTAT, Inc.

Examples

3.1 Pearson correlation

Here is an example of **Pearson...** correlations on the MEDICAL data set. Each case is a U.S. state. The matrix is computed with the following commands:

- Select **Corr/Pearson...** from the **Stats** menu
- Select the variables to be analyzed: ACCIDENT, CARDIO, CANCER, PULMONAR, PNEU_FLU, DIABETES, LIVER
- Click OK

```
Pearson Correlation Matrix

                ACCIDENT      CARDIO      CANCER    PULMONAR     PNEU_FLU

ACCIDENT           1.000
  CARDIO          -0.506       1.000
  CANCER          -0.578       0.908       1.000
PULMONAR          -0.099       0.441       0.438       1.000
PNEU_FLU          -0.196       0.538       0.358       0.400       1.000
DIABETES          -0.517       0.619       0.709       0.227       0.022
   LIVER          -0.155       0.136       0.363       0.263      -0.097

                DIABETES        LIVER

DIABETES           1.000
   LIVER           0.148       1.000

Number of Observations:   50
```

3.2 Probabilities

Let's compute the same correlations above but add the [Probabilities] option.

- Select **Corr/Pearson...** from the **Stats** menu
- Select the variables to be analyzed: ACCIDENT, CARDIO, CANCER, PULMONAR, PNEU_FLU, DIABETES, LIVER
- Click [Probabilities]
- Click OK

© 1989, SYSTAT, Inc.

```
Pearson Correlation Matrix

                ACCIDENT      CARDIO      CANCER     PULMONAR     PNEU_FLU

  ACCIDENT         1.000
    CARDIO        -0.506       1.000
    CANCER        -0.578       0.908       1.000
  PULMONAR        -0.099       0.441       0.438       1.000
  PNEU_FLU        -0.196       0.538       0.358       0.400       1.000
  DIABETES        -0.517       0.619       0.709       0.227       0.022
     LIVER        -0.155       0.136       0.363       0.263      -0.097

                DIABETES       LIVER

  DIABETES         1.000
     LIVER         0.148       1.000

 Bartlett Chi-Square Statistic:  202.766 DF=   21 PROB=  .000

Matrix of Probabilities

                ACCIDENT      CARDIO      CANCER     PULMONAR     PNEU_FLU

  ACCIDENT         0.000
    CARDIO         0.000       0.000
    CANCER         0.000       0.000       0.000
  PULMONAR         0.492       0.001       0.001       0.000
  PNEU_FLU         0.174       0.000       0.011       0.004       0.000
  DIABETES         0.000       0.000       0.000       0.113       0.878
     LIVER         0.282       0.346       0.010       0.065       0.505

                DIABETES       LIVER

  DIABETES         0.000
     LIVER         0.306       0.000

 Number of Observations:   50
```

This gives us "*p* values" for each correlation in the matrix. By themselves, these values are not very informative. An argument in their favor is that they allow one to assess the significance of the correlations. That assertion is plainly false unless there is only one probability printed. If only one correlation exists, one might as well use **MGLH** to find its significance and other information. If there are several correlations, the "*p* values" are pseudo-probabilities because they do not reflect the number of correlations being tested. If pairwise deletion is used, the problem is even worse, although many statistics packages print probabilities as if they meant something in this case too.

© 1989, SYSTAT, Inc.

In practice, these probabilities are often abused. We once saw a researcher hold up a printout containing what must have been 5,000 correlations and navigate through the matrix toward all the correlations with three stars. This practice is widespread and reflects a fundamental misunderstanding of the use of probability and statistics (see Salsburg, 1985). A good antidote is to take the SYSTAT normal random number generator, make a file of random variates, and correlate the variables in the file with **Corr/Pearson...**

SYSTAT computes the Bartlett chi-square test whenever you request probabilities for more than one correlation. This tests a global hypothesis concerning the significance of all the correlations in the matrix. If the Bartlett test is not significant, don't even look at the significance of individual correlations. In this example the test is significant, which indicates that there may be some real correlations among the variables. Keep in mind, however, that these data are aggregated over states and this presents problems in defining a population for the data and, furthermore, the Bartlett test is sensitive to nonnormality and can be used only as a rough guide.

3.3 Bonferroni-adjusted probabilities

Even if the Bartlett test is significant, you cannot accept the nominal *p* values as the true family probabilities associated with each correlation. As an alternative, examine Bonferroni-adjusted probabilities that guarantee that the family comparison error rate will not be larger than a particular critical value you select. If you are on a wild fishing expedition, these probabilities will sober you up really fast.

Here is the same example with the [Bonferroni] option:

- Select **Corr/Pearson...** from the **Stats** menu
- Select the variables to be analyzed: ACCIDENT, CARDIO, CANCER, PULMONAR, PNEU_FLU, DIABETES, LIVER
- Click [Bonferroni]
- Click OK

© 1989, SYSTAT, Inc.

Matrix of Bonferroni Probabilities

	ACCIDENT	CARDIO	CANCER	PULMONAR	PNEU_FLU
ACCIDENT	0.000				
CARDIO	0.004	0.000			
CANCER	0.000	0.000	0.000		
PULMONAR	1.000	0.028	0.031	0.000	
PNEU_FLU	1.000	0.001	0.226	0.083	0.000
DIABETES	0.003	0.000	0.000	1.000	1.000
LIVER	1.000	1.000	0.202	1.000	1.000

	DIABETES	LIVER
DIABETES	0.000	
LIVER	1.000	0.000

Number of Observations: 50

Notice that some correlations, such as the one between CANCER and PNEU_FLU, can no longer be considered significant. The Bonferroni criterion provides protection for multiple tests.

© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

4 Factor

Introduction 72
A principal component 72
Multiple principal components 74
Exploratory vs. confirmatory components 75
Component coefficients 76
Component loadings 76
Correlations or covariances? 76
Rotation 77
Signs of component loadings 77
Factor analysis 78
Estimating factors 79
Principal components vs. factor analysis 79
Applications and caveats 82

Usage 83
Principal components... 83
Type of matrix to factor 84
Pairwise deletion of missing data 84
Number of factors and Eigenvalue 85
Rotation 85
Plot factors 85
Labels 85
Sort loadings 85
Save file 85
Show complete results 86
Common factor model... 86
Iterations 86
Tolerance 86
Save file 86

Computation 87
Missing data 87
Algorithms 87

Examples 88
4.1 Principal components 88
4.2 Plot component scores 90

© 1989, SYSTAT, Inc.

Factor 4

Overview

Factor produces either a principal components analysis or an iterated principal axis factor analysis with optional orthogonal rotations. Scores on principal components or rotated components can be saved into a SYSTAT file for processing by items in **Stats** and **Graph**. For example, you can obtain principal components for a set of data, compute scores, and use **MGLH** to regress one or more variables onto the components you computed with **Factor**.

MGLH offers a principal components analysis as well, but it doesn't allow you to save scores. Use **MGLH** only to factor within-groups or residual covariance or correlation matrices, since ordinary principal components and rotations are easier to do with **Factor**.

© 1989, SYSTAT, Inc.

Introduction

Principal component and factor analyses are methods of decomposing a correlation or covariance matrix. Although they are based on different mathematical models, both can be used on the same data and both usually produce similar results. If you understand the component model, you are on the way toward understanding the factor model, so let's begin with the former.

A principal component

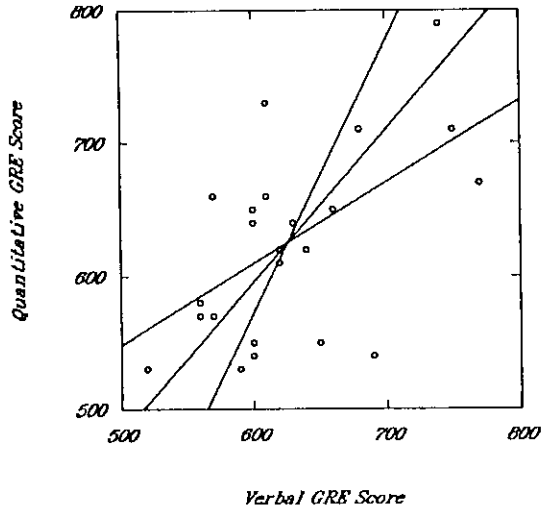
What is a principal component? The simplest way to see is through real data. The following data consist of Graduate Record Examination verbal and quantitative scores. These scores are from 25 applicants to a graduate psychology department.

VERBAL	QUANTITATIVE
590	530
620	620
640	620
650	550
620	610
610	660
560	570
610	730
600	650
740	790
560	580
680	710
600	540
520	530
660	650
750	710
630	640
570	660
600	650
570	570
600	550
690	540
770	670
610	660
600	640

© 1989, SYSTAT, Inc.

Now, we could decide to predict verbal scores from quantitative. The introduction to **MGLH** gives you the background of how to do this with regression modeling. Or, we could decide to predict quantitative from verbal by the same method. The data don't suggest which is a dependent variable; either will do. What if we aren't interested in predicting either one separately but instead want to know how both variables hang together jointly? This is what a principal component does. Karl Pearson, who developed principal component analysis in 1901, described a component as a "line of closest fit to systems of points in space." In short, the regression line indicates best *prediction* and the component line indicates best *association*.

The following figure shows the regression and component lines for our GRE data. The regression of y on x is the line with the smallest slope (flatter than diagonal). The regression of x on y is the line with the largest slope (steeper than diagonal). The component line is between the other two. Interestingly, when most people are asked to draw a line relating two variables in a scatterplot, they tend to approximate the component line. It takes much explaining to get them to realize that this is not the best line for *predicting* the vertical axis variable (y) or the horizontal axis variable (x).



© 1989, SYSTAT, Inc.

Notice that the slope of the component line is approximately 1, which means that the two variables are weighted almost equally (assuming the axis scales are the same). We could make a new variable called GRE that is the sum of the two tests:

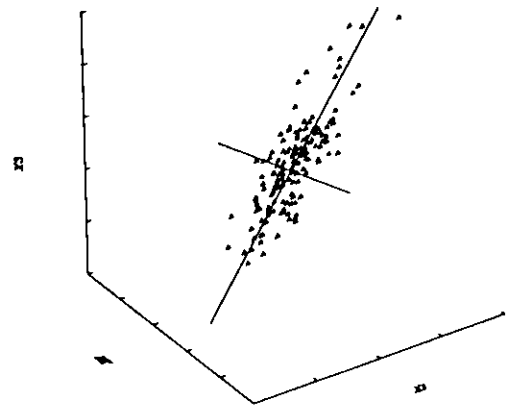
GRE = VERBAL + QUANTITATIVE

This new variable could summarize, albeit crudely, the information in the other two. If the points clustered almost perfectly around the component line, then the new component variable could summarize almost perfectly both variables.

Multiple principal components

The goal of principal components analysis is to summarize a multivariate dataset as accurately as possible using a few components. So far, we have seen only one component. It is possible, however, to draw a second component perpendicular to the first. The first component will summarize as much of the joint variation as possible. The second will summarize what's left. If we do this with the GRE data, of course, we will have as many components as original variables—not much of a saving. We usually seek fewer components than variables, so that the variation left over is negligible. Here is an example in which three variables (X1, X2 and X3) are summarized almost completely by two principal components. We have drawn the component axes in perspective. The first principal component has the longest axis. The second is next longest, perpendicular to the first. The third is next longest, perpendicular to the first and second. In this example, the third component axis is not visible because it is negligible.

© 1989, SYSTAT, Inc.



Exploratory vs. confirmatory components

Notice that we have ended up with two sets of axes: the original variables axes and the components axes. Either set can describe the data. Often, we give the component axes names, as in our Graduate Record Examination example, where we called the first component GRE. If you name components by referring to the meaning of the original variables or their names, which is the usual procedure, keep in mind that this reasoning is circular. For some researchers, components (and factors) take on a life of their own, as if they are something floating out in the ideal world waiting to be discovered. If you use a computer program to compute components on your data, you are searching for quantitative associations between variables—nothing more, nothing less.

There is nothing wrong with the search for associations between variables during exploratory phases of research in a new field. At this point, discovering that variables cluster together can lead to new hypotheses. A second stage of research would involve searching for similar components in a new sample of data. A third stage would be to test the structure of hypothesized components using a confirmatory program like LISREL, EQS, or EzPATH, a supplementary module for SYSTAT.

Another use for components as measures of association is to construct test scales made up of multiple questions. Usually, items that measure a common dimension should be associated in a components analysis. You can weed out irrelevant items this way. An alternative procedure for constructing test scales is called *latent trait analysis*, which is offered in TESTAT, a supplementary module for SYSTAT.

© 1989, SYSTAT, Inc.

Component coefficients

In the above equation for computing the first principal component on our test data, we made both coefficients equal. In fact, when you run the sample covariance matrix through the FACTOR program, the coefficients are as follows:

GRE = .008*VERBAL + .01*QUANTITATIVE

They are indeed nearly equal. Their magnitude is considerably less than one because principal components are usually scaled to *conserve variance*. That is, once you compute the components with these coefficients, the total variance on the components is the same as the total variance on the original variables.

Component loadings

Most researchers want to know the relation between the original variables and the components. Some components may be nearly identical to an original variable; in other words, their coefficients may be nearly zero for all variables except one. Other components may be a more even amalgam of several original variables. *Component loadings* are the covariances of the original variables with the components. In our example, these loadings are 51.085 for Verbal and 62.880 for Quantitative. You may have noticed that these are proportional to the coefficients; they are simply scaled differently. If you square each of these loadings and add them up separately for each component, you will have the variance accounted for by each component.

Correlations or covariances?

Some researchers prefer to analyze the correlation rather than covariance structure among their variables. Sample correlations are simply covariances of sample standardized variables. Thus, if you feel the standard deviations of your variables are not theoretically significant, you may wish to work with correlations instead of covariances. In our test example, working with correlations yields loadings of .879 for each variable instead of 51.085 and 62.880. When you factor the correlation instead of covariance matrix, then the loadings are the correlations of each component with each original variable.

For our test data, loadings of .879 mean that if you created a GRE component by standardizing Verbal and Quantitative and adding them together weighted by the coefficients, you would find the correlation between these component scores and the original Verbal scores to be .879. The same would be true for Quantitative.

© 1989, SYSTAT, Inc.

Rotation

Components are uniquely defined mathematically. Each is perpendicular to the others in the component space and each accounts for a maximum portion of the available variance. They are not laws of nature, however. Components will not necessarily be more theoretically meaningful than any other linear combination of your variables. Consequently, many researchers prefer to use components to reduce dimensionality by eliminating negligible variation and then rotate *once more* in the smaller dimensional space to achieve some meaningful criterion. You can use the SYSTAT rotation options to do this in three dimensions.

More generally, researchers let the computer do the rotation automatically. There are many criteria for achieving a *simple structure* among component loadings, although Thurstone's are most widely cited. For p variables and m components:

1) Each component should have at least m near-zero loadings.
2) Few components should have nonzero loadings on the same variable.

SYSTAT offers several methods for approximating these conditions. The most popular is the default [Varimax] rotation. We won't show a rotation for our test data, because rotating components based on fewer than three variables is pointless. Take a look at the rotated MEDICAL data later in this chapter for an example.

Signs of component loadings

The signs of loadings within components are arbitrary. If a component (or factor) has more negative than positive loadings, you may change minus signs to plus and plus to minus. This merely reverses the polarity of a component. It's like rotating a component or factor 180°. SYSTAT does this automatically for components that have more negative than positive loadings, and thus will occasionally produce components or factors that have different signs from those in other computer programs. This occasionally confuses users. At least one publication (Hubbard and Allen, 1987) has committed the error of regarding relative signs of loadings as theoretically significant rather than artifactual. In mathematical terms $Ax=\lambda x$ and $-Ax=-\lambda x$ are equivalent.

© 1989, SYSTAT, Inc.

Factor analysis

We have seen how principal components analysis is a method for computing new variables that summarize variation in a space parsimoniously. For our test variables, the equation for computing the first component was:

GRE = .008*VERBAL + .01*QUANTITATIVE

This component equation is linear, of the form:

Component = Linear combination of (Observed variables)

Factor analysts turn this equation around:

Observed variable = Linear combination of (Factors) + Error

This model was presented by Spearman near the turn of the century in the context of a single intelligence factor and extended to multiple mental measurement factors by Thurstone several decades later. Notice that the factor model makes observed variables a function of unobserved factors. Even though this looks like a linear regression model, none of the graphical and analytical techniques used for regression can be applied to the factor model because there is no unique, observable set of factor scores or residuals to examine.

Factor analysts are less interested in prediction than in decomposing a covariance matrix. This is why the fundamental equation of factor analysis is not the above linear model, but rather its *quadratic form*:

Observed covariances = Factor covariances + Error covariances

The covariances in this equation are usually expressed in matrix form, so that the model decomposes an observed covariance matrix into a hypothetical factor covariance matrix plus a hypothetical error covariance matrix. The diagonals of these two hypothetical matrices are known, respectively, as *communalities* and *specificities*. In ordinary language, then, the factor model expresses variation within and relations among observed variables as partly common variation among factors and partly specific variation among random errors.

© 1989, SYSTAT, Inc.

Estimating factors

Factors must be estimated iteratively in a computer. There are several methods available. The most popular approach, used in SYSTAT, is to modify the diagonal of the observed covariance matrix and calculate factors the same way components are computed. This procedure is repeated until the communalities reproduced by the factor covariances are indistinguishable from the diagonal of the modified matrix.

An alternative is to use the method of maximum likelihood. This technique is beyond the scope of this introduction, but is offered in the SYSTAT supplement EzPATH. EzPATH also analyzes confirmatory factor models, path models, and causal models.

Principal components vs. factor analysis

Factor can perform both principal components and common factor analysis. Some view principal components analysis as a method of factor analysis, though there is a theoretical distinction between the two. Principal components are weighted linear composites of observed variables. Common factors are unobserved variables that are hypothesized to account for the intercorrelations among observed variables.

One significant practical difference is that common factor scores are indeterminate, whereas principal component scores are not. There are no sufficient estimators of scores for subjects on common factors (rotated or unrotated, maximum likelihood or otherwise). Some computer models provide "regression" estimates of factor scores, but these are not estimates in the usual statistical sense. This problem arises not because factors can be arbitrarily rotated (so can principal components), but because the common factor model is based on more unobserved parameters than observed data points, an unusual circumstance in statistics.

In recent years, "maximum likelihood" factor analysis algorithms have been devised to estimate common factors. The implementation of these algorithms in popular computer packages has led some users to believe that the factor indeterminacy problem does not exist for "maximum likelihood" factor estimates. It does.

Statisticians unfamiliar with the common factor model may recognize an analogous controversy regarding a related procedure: ridge regression. Ridge regression "inflates" dimensionality, as the common factor model "deflates" dimensionality, by altering the diagonal of a sample covariance matrix. In doing so, both methods relinquish the functional relationship between the raw data and the covariance matrix in order to gain parsimony.

© 1989, SYSTAT, Inc.

Mathematicians and psychometricians have known about the factor indeterminacy problem for decades. For a historical review of the issues see Steiger (1979), and for a general review, see Rozeboom (1982). For further information on principal components, consult Harman (1976), Mulaik (1978), Gnanadesikan (1977), or Mardia, Kent, & Bibby (1979).

Because of the indeterminacy problem, Factor computes subjects' scores only for the principal components model where subjects' scores are a simple linear transformation of scores on the factored variables. Factor does not save scores from a common factor model.

Although some factor analysts become apoplectic over these issues and wish to reserve the term "factor analysis" for the common factor model, principal component and common factor solutions for real data rarely differ enough to matter. After I made this statement in an earlier SYSTAT manual, Borgatta, Kercher, and Stull (1986) contrived an artificial data example to show that common factor and principal component models can indeed differ (I never said they couldn't). In doing so, they unwittingly provided a counter-example that reveals the fundamental indeterminacy in the factor model.

Here is the matrix they used to illustrate their point. The entries 0.35355 are in fact $\sqrt{2}/4$.

	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7
VAR1	1.00000						
VAR2	0.50000	1.00000					
VAR3	0.35355	0.35355	1.00000				
VAR4	0.00000	0.00000	0.35355	1.00000			
VAR5	0.00000	0.00000	0.35355	0.50000	1.00000		
VAR6	0.00000	0.00000	0.25000	0.35355	0.35355	1.00000	
VAR7	0.00000	0.00000	0.00000	0.00000	0.00000	0.35355	1.00000

I invite you to factor this matrix into three common factors with SYSTAT and other reliable mainframe and microcomputer statistics packages (e.g. SAS, BMDP, SPSS). If you want, you can Varimax rotate the factors as Borgatta, Kercher and Stull did. You will find, as I did, that every program gives a different result depending on the initial communalities chosen. Since the authors used only SPSS, they did not detect this anomaly. Some loadings differ between packages by as much as .09, which is greater than some researchers' distinctions in practice. Are the programs wrong? No. I verified that the SAS, BMDP, SYSTAT, and SPSS results are all valid. They all reproduce the correlation matrix exactly.

© 1989, SYSTAT, Inc.

How can all answers be correct? The explanation is that this correlation matrix is under-identified. This anomaly has nothing to do with the arbitrariness of rotations for multiple factors. Even the unrotated factor solutions differ for this matrix and if you try Procrustes rotations to match them, you will find they do not fit. They really are different. On the other hand, there is only *one* principal components solution to this matrix.

To quote from their paper, "Locating a parsimonious set of factors that will reproduce the correlation matrix is the homely task with which most social scientists are concerned." In this instance, their correlation matrix can be reproduced by an *infinite* number of different common factors, so the task is not homely; it is arbitrary. I do not think Borgatta, Kercher, and Stull speak for a majority of social scientists in grounding social science methodology on their "homely task." Most social scientists I know focus on falsifiability of theories using observed data rather than on arbitrary decompositions of correlation matrices. Borgatta, Kercher, and Stull may feel that the common factor model can elucidate theories of social science, but I believe that a mathematical method that can produce multiple and sometimes contradictory solutions on the *same* data is dangerous.

Borgatta, Kercher, and Stull described me, *ad hominem*, as "a person who has not used factor-analytic procedures extensively in a great variety of circumstances." Seventeen years ago, I published one of the first applications of the maximum likelihood common factor model (Wilkinson, 1973) and subsequently factored a lot of data before losing faith in the method. If you think real data cannot produce conflicting results on different common factor programs the way the above example did, then you have not had "extensive" experience with the common factor model. After "extensive" experience with the model, my advice is "beware." Stick with principal components.

Applications and caveats

You may want to analyze the MEDICAL statistics example later in this chapter using the factor model in SYSTAT instead of the component model. The results are quite similar. While there is not room here to discuss more statistical issues, you should realize that there are several myths about factors vs. components:

© 1989, SYSTAT, Inc.

1) **MYTH: the factor model allows hypothesis testing. The component model doesn't.** Morrison (1967) and others present a full range of formal statistical tests for components.

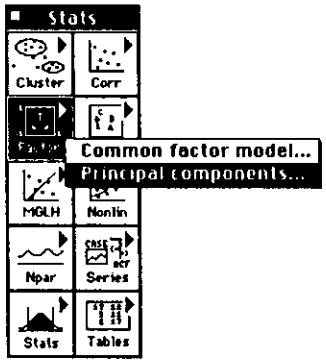
2) **MYTH: factor loadings are real. Principal component loadings are approximations.** This statement is too ambiguous to have any meaning. It is easy to define things so that factors are approximations of components.

3) **MYTH: factor analysis is more likely to uncover lawful structure in your data. Principal components are more contaminated by error.** Again, this statement is ambiguous. With further definition, it can be shown to be true for some data, false for other. It is true that, in general, factor solutions will have lower dimensionality than corresponding component solutions. This can be an advantage to you when searching for simple structure among noisy variables, as long as you compare the result to a principal components solution to avoid being fooled by the sort of degeneracies illustrated above.

© 1989, SYSTAT, Inc.

Usage

Factor provides both principal components analysis and (iterated principal axis) common factor analysis. With a rectangular data file, **Factor** computes a correlation or covariance matrix before factoring. Otherwise, it will accept a correlation or covariance or SSCP matrix directly from a SYSTAT file.



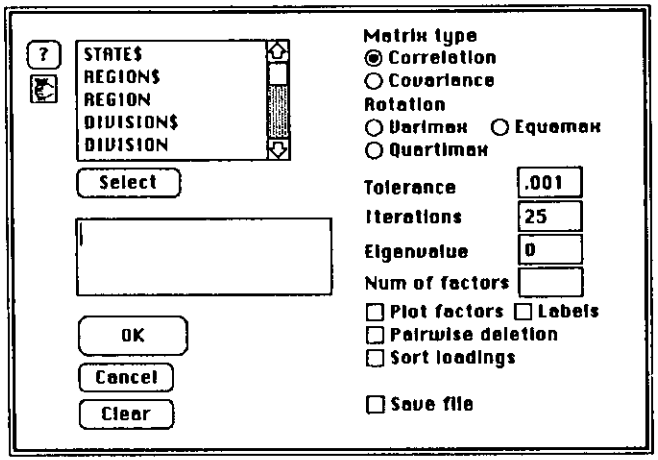
Press and hold the mouse button on **Factor**. A submenu pops up offering **Common factor model...** and **Principal components...** Select the item you want and release the mouse button.

Principal components...

Principal components... offers principal components analysis.

- Select **Factor/Principal components...** from the **Stats** menu
- Select variables to be analyzed; all numeric variables are used by default
- Click OK

© 1989, SYSTAT, Inc.



Type of matrix to factor

If you use a rectangular data file, you may choose to factor the [Correlation] or [Covariance] matrix. If you use a symmetric SYSTAT file, the type of matrix is read from the file and cannot be changed.

Pairwise deletion of missing data

Ordinarily, Factor and other multivariate procedures in SYSTAT delete all cases having missing values on any variable selected for analysis. This is called *listwise deletion*. For data with many missing values, you may end up with too few cases for analysis. Click [Pairwise deletion] if you want covariances or correlations to be computed separately for each pair of variables selected for analysis. Pairwise deletion takes considerably more computer time than the standard listwise deletion, since all possible pairs of variances and covariances must be computed. The same option is offered in **Corr**, should you decide to create a symmetric matrix for use in **Factor** that way.

Be careful. When you use the [Pairwise deletion] option, you can end up with negative eigenvalues for principal components or be unable to compute common factors at all. With either method, it is desirable that the pattern of missing data be random. Otherwise, the factor structure you compute will be influenced systematically by the pattern of missing data.

© 1989, SYSTAT, Inc.

Number of factors and Eigenvalue

Two options allow you to limit the number of components or factors computed. With the [Eigenvalue] option, you can specify the smallest eigenvalue to be retained. The default is 0. Incidentally, the standard value of zero causes **Factor** to ignore components with negative eigenvalues (which can be produced with pairwise deletion). The [Number of factors] option lets you limit the number of factors directly. If you use both limiting options, **Factor** uses whichever criterion results in the smaller number of components.

Rotation

Factor offers three types of rotation: [Varimax], [Equamax], and [Quartimax]. See Harman (1976) for an explanation of these rotation methods. If you request rotation, SYSTAT prints the rotated loadings as well as the sum of squared rotated loadings (variance accounted for by each factor).

Plot factors

Check the [Plot factors] box if you wish to plot loadings for all possible pairs of factors. This can take a lot of space if you have a large number of factors, so the default setting is not to plot.

Labels

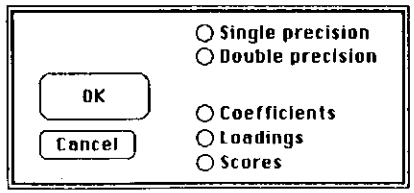
Check the [Labels] box if you wish to label each loading with its respective variable name.

Sort loadings

If you click the [Sort loadings] option, **Factor** sorts the rows of the loadings matrix so that high loadings (greater than .5) on each component are adjacent and list from highest to lowest. If the loadings matrix is rotated, **Factor** again sorts the loadings on the rotated matrix to this pattern. The resulting sort makes it easier to discern the "simple structure" defined by Thurstone (see the "Introduction").

Save file

You can save component scores, loadings, or coefficients in a SYSTAT file by checking the [Save file] box. A dialog box lets you choose what to save: [Coefficients], [Loadings], or [Scores]. Loadings and coefficients are for unrotated components or factors unless you rotate before factoring.



© 1989, SYSTAT, Inc.

If you save scores, the variables in the file are labeled FACTOR(1), FACTOR(2), etc. Any observations with missing values on any of the input variables will have missing values on all of the scores. The scores are normalized to have zero mean and, if the correlation matrix is used, unit variance.

If you use the covariance matrix and perform no rotations, SYSTAT does not standardize the component scores. The sum of their variances is the same as for the original data.

Show complete results

If you choose [Show extended results] in **Data/Formats...**, SYSTAT displays the correlation or covariance matrix you are factoring.

Common factor model...

Common factor... offers common factor analysis.

- Select **Factor/Common factor model...** from the **Stats** menu
- Select variables to be analyzed; all numeric variables are used by default
- Click OK

All the options for **Factor/Principal components...** work for **Factor/Common factor model...**. In addition, the [Tolerance] and [Iterations] boxes are enabled. The [Save file] box has minor differences as well, so it will be reviewed here.

Iterations

The [Iterations] option specifies the maximum number of iterations used to compute communalities. The default is 25. If you fail to get convergence, you can increase this number.

Tolerance

[Tolerance] specifies the convergence criterion (change in communality estimates) for stopping iterations. The default value is 0.001.

Save file

You can save factor loadings into a SYSTAT file by checking the [Save file] box. You cannot save factor scores or coefficients because they cannot be uniquely determined.

© 1989, SYSTAT, Inc.

Computation

Missing data

Any case with missing data on any variable selected in the **Factor** command is excluded from computations. If the [Pairwise] option is clicked, pairwise deletion of missing data is used instead. Component scores are set to missing values for any cases missing any data on any variables included in the analysis.

Algorithms

Factor is programmed in double precision. Provisional methods are used for computing covariance or correlation matrices (see the **Corr** chapter for references). Components are computed via a Householder tridiagonalization and implicit QL iterations (see the **MGLH** chapter for references). Rotations are computed with a variant of Kaiser's iterative algorithm, described in Mulaik (1972).

The factor model uses squared multiple correlations as initial estimates of communalities. If the squared multiple correlations cannot be computed (e.g. a singular correlation matrix) SYSTAT uses maximum absolute correlations as initial communality estimates.

© 1989, SYSTAT, Inc.

4.1 Principal components

This example uses the MEDICAL variables (deaths-by-cause). The cases are U.S. states. These same data are used in a principal components analysis in the **MGLH** chapter with regional effects removed. In this analysis, we ignore the DIVISION variable to see if the total structure is similar.

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Click OK

- Select **Factor/Principal components...** from the **Stats** menu
- Select variables to be analyzed: ACCIDENT, CARDIO, CANCER, PULMONAR, PNEU_FLU, DIABETES, LIVER
- Choose [Varimax] rotation
- Specify number of components: 2
- Click [Plot factors]
- Click [Sort matrix]
- Click [Save file]; save a file of [Component scores]
- Click OK

Matrix to be Factored

	ACCIDENT	CARDIO	CANCER	PULMONAR	PNEU_FLU
ACCIDENT	1.000				
CARDIO	-0.506	1.000			
CANCER	-0.578	0.908	1.000		
PULMONAR	-0.099	0.441	0.438	1.000	
PNEU_FLU	-0.196	0.538	0.358	0.400	1.000
DIABETES	-0.517	0.619	0.709	0.227	0.022
LIVER	-0.155	0.136	0.363	0.263	-0.097

	DIABETES	LIVER
DIABETES	1.000	
LIVER	0.148	1.000

Latent Roots (Eigenvalues)

1	2	3	4	5
3.401	1.229	1.061	0.611	0.434

6	7
0.216	0.047

© 1989, SYSTAT, Inc.

Component Loadings

	1	2
CANCER	0.951	-0.109
CARDIO	0.921	0.148
DIABETES	0.739	-0.410
ACCIDENT	-0.666	0.310
PULMONAR	0.558	0.417
PNEU_FLU	0.489	0.771
LIVER	0.326	-0.405

Variance Explained by Components

1	2
3.401	1.229

Percent of Total Variance Explained

1	2
48.587	17.561

Rotated Loadings

	1	2
DIABETES	0.833	0.146
CANCER	0.808	0.513
ACCIDENT	-0.713	-0.178
CARDIO	0.624	0.694
LIVER	0.508	-0.110
PNEU_FLU	-0.104	0.907
PULMONAR	0.172	0.675

Variance Explained by Rotated Components

1	2
2.543	2.087

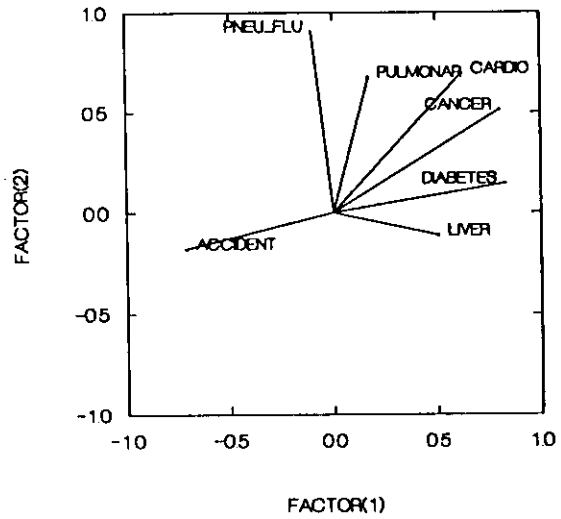
Percent of Total Variance Explained

1	2
36.328	29.820

Factor Score Coefficients

	1	2
DIABETES	0.379	-0.123
CANCER	0.273	0.107
ACCIDENT	-0.311	0.073
CARDIO	0.135	0.026
LIVER	0.281	-0.196
PNEU_FLU	-0.282	0.578
PULMONAR	-0.086	0.367

© 1989, SYSTAT, Inc.



If you want to use the score coefficients to get component scores for new data, you must multiply the coefficients by the standardized data. SYSTAT automatically does this when it saves scores. Another way to do cross-validation would be to assign a zero weight on a weighting variable to cases not included in the factoring and unit weight to cases to be included. The zero weight cases would not be used in the factoring, but scores would be computed for all cases.

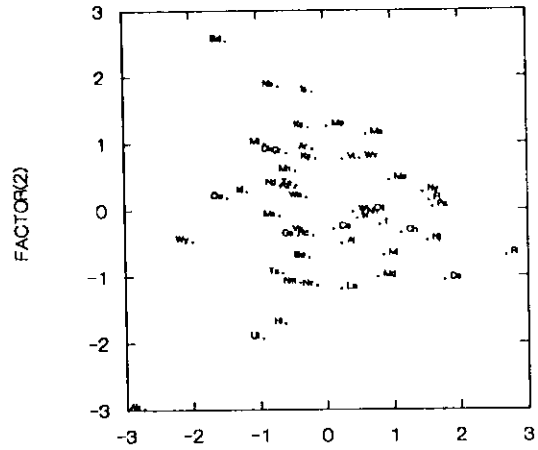
4.2 Plot component scores

To show the variation among states on these medical components, we can plot the rotated component scores from the analysis. In the previous example, we saved the scores into the file MEDICAL Factor. To plot the scores, we have added labels for the states in the editor (or, you could merge the original datafile labels). Now, we can plot the scores.

- Select **Plot/Plot** from the **Graph** menu
- Select the Y variable: FACTOR(2)
- Select the X variable: FACTOR(1)
- Click [Symbol] and specify STATE$ in the box next to the symbol choices
- Click OK

© 1989, SYSTAT, Inc.

Notice that urban states score positively and rural negatively on factor 1. The diseases associated with this factor are diabetes and cancer and low accident rates. Factor 2 is more ambiguous, but it appears to represent states with high incidences of lung (and heart) diseases. These tend to be Midwest or cold and damp states, but there are notable exceptions like Alaska (in the lower left corner).



© 1989, SYSTAT, Inc.

5 MDS

Introduction 94
Assumptions 94
Collecting dissimilarity data 95
Direct methods 95
Indirect methods 96
Scaling dissimilarities 97
Stress and iterations 97
The Shepard diagram 97
The MDS plot 97
An example 98

Usage 100
Scale... 100
Controlling convergence 101
Number of dimensions 101
Choosing a distance metric 102
Shepard plots 102
Choosing a loss function 102
Choosing a regression function 102
Saving files 102
Configuration... 103
Confirmatory scaling 104

Computation 105
Missing data 105
Algorithms 105
Kruskal method 105
Guttman method 106

Examples 108
5.1 Guttman method 108
5.2 Kruskal method 111

© 1989, SYSTAT, Inc.

MDS 5

Overview

MDS offers nonmetric multidimensional scaling of a similarity or dissimilarity matrix in 1 to 5 dimensions. Multidimensional scaling is a powerful data reduction procedure that can be used on a direct similarity or dissimilarity matrix or on one derived from rectangular data with Corr. Although there are various multidimensional scaling algorithms, all perform a similar function: to compute coordinates for a set of points in a space such that the distances between pairs of these points fit as closely as possible to measured dissimilarities between a corresponding set of objects.

The family of procedures called principal components or factor analysis is related to multidimensional scaling in function, but multidimensional scaling differs from this family in important respects. Usually, but not necessarily, multidimensional scaling can fit an appropriate model in fewer dimensions than can these other procedures. Furthermore, if it is implausible to assume a linear relationship between distances and dissimilarities, multidimensional scaling nevertheless provides a simple dimensional model. For more information see Borg (1981), Carroll and Arabie (1980), Green and Rao (1972), Schiffman, Reynolds, and Young (1981), and Shepard, Romney, and Nerlove (1972).

© 1989, SYSTAT, Inc.

Introduction

Multidimensional scaling (**MDS**) is a procedure for fitting a set of points in a space such that the distances between points correspond as closely as possible to a given set of dissimilarities between a set of objects. Dissimilarities may be measured directly, as in psychological judgments, or derived indirectly, as in correlation matrices computed on rectangular data.

Assumptions

Because **MDS**, like cluster analysis, operates directly on dissimilarities, there are no statistical distribution assumptions necessary. There are other important assumptions, however. First, multidimensional scaling is a spatial model. To fit points in the kinds of spaces **MDS** covers, you must assume that your data satisfy *metric* conditions. These are:

1) The distance from an object to itself is zero.
2) The distance from object A to object B is the same as that from B to A.
3) The distance from object A to C is less than or equal to the distance from A to B plus B to C. This is sometimes called the *triangle inequality*.

You may think these conditions are obvious, but there are numerous counter-examples in psychological perception and elsewhere. For example, commuters often view the distance from home to the city as closer than the distance from city to home because of traffic patterns, terrain, and psychological expectations related to time of day. Framing or context effects can also disrupt the metric axioms, as Amos Tversky has shown. For example, Miami is similar to Havana. Havana is similar to Moscow. Is Miami similar to Moscow? If your data (objects) are not consistent with these three axioms, don't use **MDS**.

Second, there are ways of deriving distances from rectangular data that do not satisfy the metric axioms. The ones available in **Corr** do, but if you are thinking of cooking up some other derived measure of similarity, check it carefully.

© 1989, SYSTAT, Inc.

Finally, it is assumed that all your objects will fit in the same metric space. It is best if they diffuse somewhat evenly through this space as well. Don't expect to get interpretable results for 25 nearly indistinguishable objects and one radically different

Collecting dissimilarity data

You may collect dissimilarities directly or compute them indirectly. Examples of direct dissimilarities are:

Direct methods

Distances

Take distances between objects (e.g. cities) directly off a map. If the scale is local, **MDS** will reproduce the map nicely. If the scale is global, you will need three dimensions for a **MDS** fit. Two or three dimensional spatial distances can be measured directly. Direct measures of social distance might include spatial propinquity or the number of times or amount of time one individual interacts with another.

Judgments

Ask subjects to give a numerical rating of the dissimilarity (e.g. 0 to 10) between all pairs of objects.

Clusters

Ask people to sort objects into piles. Or examine naturally occurring aggregates, such as paragraphs, communities, associations. Record 0 if two objects occur in the same group, 1 if they do not. Sum these counts over replications or judges.

Triads

Ask subjects to compare three objects at a time and report which two are most similar (or which is odd one out). Do this over all possible triads of objects. To compute dissimilarities, sum over all triads, as for the clustering method. There are usually many more triads than pairs of objects, so this method is more tedious. It allows you to assess independently possible violations of the triangle inequality, however.

© 1989, SYSTAT, Inc.

Indirect methods

Indirect dissimilarities are computed over a rectangular matrix whose columns are the objects and rows are attributes You can transpose this matrix if you wish to scale rows instead. Possible indirect dissimilarities are:

Computed Euclidean distances
These are the square root of the sum of squared discrepancies between columns of the rectangular matrix.

Negatives of correlations
For standardized data (mean zero and standard deviation one) Pearson correlations are proportional to Euclidean distances. For unstandardized data, Pearson correlations are comparable to computing Euclidean distances after standardizing. **MDS** automatically negates correlations if you do not. Other types of correlations—e.g. Spearman, Gamma—are analogous to standardized distances, but only approximately. Also, be aware that large negative correlations will be treated as large distances and large positive correlations as small distances. Make sure all variables are scored in the same direction before computing correlations. If you find a whole row of a correlation matrix negative, reverse the variable by multiplying by –1 and recompute the correlations.

Counts of discrepancies
Counting discrepancies between columns or using some of the binary association measures in **Corr** is closely related to computing Euclidean distance. These methods are also related to the clustering distance calculations mentioned above for direct distances.

© 1989, SYSTAT, Inc.

Scaling dissimilarities

Once you have dissimilarities (or similarities, correlations, etc., which **MDS** automatically transforms to dissimilarities), you may scale them. You do not need to know how the computer does the calculations in order to use the program intelligently as long as you pay attention to the following:

Stress and iterations

Stress is the goodness of fit statistic **MDS** tries to minimize. It consists of the square root of the normalized squared discrepancies between interpoint distances in the **MDS** plot and the smoothed distances predicted from the dissimilarities. Stress varies between 0 and 1, with values near 0 indicating better fit. It is printed for each *iteration*, which is one movement of all the points in the plot toward a better solution. Make sure iterations proceed smoothly to a minimum. This is true for the examples in this chapter. If you find that the stress values *increase* or decrease in uneven steps, you should be suspicious.

The Shepard diagram

The Shepard diagram is a scatterplot of the distances between points in the **MDS** plot against the observed dissimilarities (or similarities). The points in the plot should adhere cleanly to a curve or straight line (which would be the smoothed distances). In other words, you should look at a good Shepard plot and think it resembles the outcome of a well designed experiment. Check the examples in this chapter.

If the Shepard diagram resembles a stepwise or L-shaped function, beware. You may have achieved a degenerate solution. Publish it and you will be excoriated by the clergy.

The **MDS** plot

The plot of points is what you seek. The points should be scattered fairly evenly through the space. If you are scaling in more than two dimensions, you will have to examine plots of pairs of axes, rotate it in three dimensions in SYGRAPH (either with **Spin** or a perspective plot drawn with **3-D Plot**). The orientation of axes is arbitrary—remember we are scaling distances, not axes. Feel free to reverse axes or rotate the solution. **MDS** rotates it to the largest dimensions of variation, but these don't necessarily mean anything for your data.

You may interpret the axes as in principal components or factor analysis. More often, however, you should look for clusters of objects or regular patterns among the objects, such as circles, curved manifolds, and other structures. See the colors example later in this chapter for a good view of a circle.

© 1989, SYSTAT, Inc.

An example

Let's use the cars data from the **Cluster** and **Corr** chapters. We will start by standardizing the variables (acceleration, speed, etc.), transposing the matrix, and computing Euclidean distances. Here is the result:

Matrix of Normalized Euclidean distances

	PORSCHE	FERRARI	CORVETTE	MERCEDES	SAAB
PORSCHE	0.000				
FERRARI	0.633	0.000			
CORVETTE	0.373	0.709	0.000		
MERCEDES	1.123	1.408	1.386	0.000	
SAAB	1.116	1.589	1.218	0.769	0.000
TOYOTA	0.804	1.238	0.766	1.048	0.592
BMW	0.974	1.369	1.077	0.666	0.392
HONDA	2.088	2.495	1.920	2.236	1.513
ACURA	1.292	1.625	1.037	1.740	1.156
VW	1.948	2.375	1.923	1.581	0.937
CHEVY	1.971	2.367	1.819	2.021	1.436

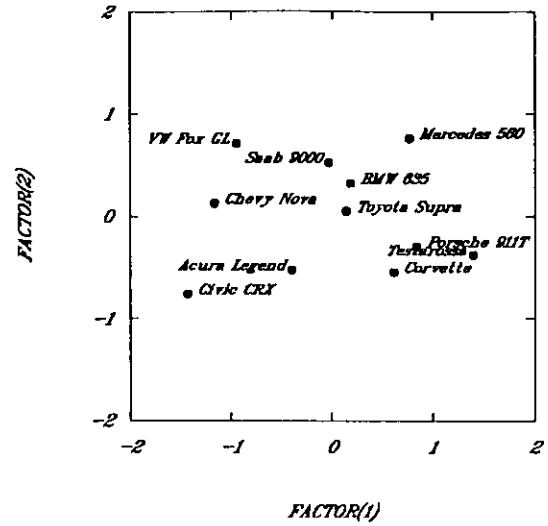
	TOYOTA	BMW	HONDA	ACURA	VW
TOYOTA	0.000				
BMW	0.466	0.000			
HONDA	1.513	1.742	0.000		
ACURA	0.732	1.124	1.128	0.000	
VW	1.193	1.125	1.161	1.276	0.000
CHEVY	1.285	1.397	1.483	1.038	1.115

	CHEVY
CHEVY	0.000

Number of observations: 5

This matrix was fed to **MDS** and the following plot was produced. We are not presenting the Shepard diagram, because it is almost a perfect straight line relating dissimilarities to distances. The stress fell smoothly to .058, which is a relatively low value. Some have given rules of thumb for good or bad stress values, but more important than absolute size of this statistic is the appearance of the Shepard diagram and the interpretability and replicability (on new data) of the solution.

© 1989, SYSTAT, Inc.

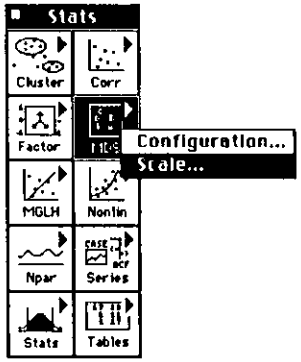


Notice that the supercars tend to appear in the lower right corner, the luxury cars in the upper right, the economy cars in the upper left, and the sporty economy cars in the lower left.

© 1989, SYSTAT, Inc.

MDS scales similarity or dissimilarity matrices. First, open a SYSTAT file containing an SSCP, covariance, correlation, or dissimilarity matrix. **MDS** automatically recognizes its type. **MDS** ignores missing values in calculating its loss function (see "Missing data" in the "Computation" section).

MDS has two items. **Scale...** does Guttman and Kruskal scaling. You can use **Configuration...** first to change from two-dimensional scaling and to specify your own starting configuration.



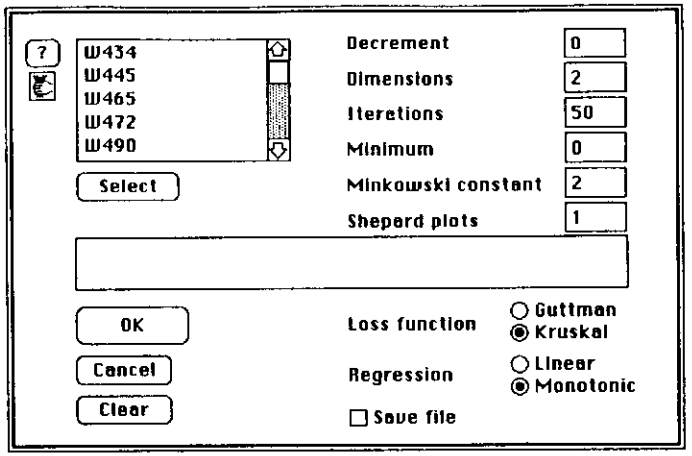
Press and hold the mouse button on **MDS**. A submenu pops up offering **Configuration...** and **Scale...** Select the item you want and release the mouse button.

Scale...

The **Scale...** dialog box provides Kruskal-Shepard-Torgerson-Young Multidimensional Scaling or Guttman Smallest Space Analysis.

- Select **MDS/Scale...** from the **Stats** menu
- Select the objects to be scaled; all objects in the matrix are used by default
- Click OK

© 1989, SYSTAT, Inc.



If you specify a subset of objects to be used, SYSTAT picks the correct rows and columns of the matrix. You do not need to create another matrix or file.

Controlling convergence

There are several methods for controlling the number of iterations needed to produce a satisfactory solution.

[Decrement]
Iterations terminate when the value of Stress differs from the value on the previous iteration by less than [Decrement]. The default is 0.

[Iterations]
Iterations terminate when the number of iterations exceeds [Iterations]. The default is 50.

[Minimum]
Iterations terminate when the value of Stress is less than [Minimum]. The default is 0.

Number of dimensions

You can specify the number of dimensions (1 through 5) in which to scale your points with the [Dimensions] box. The default is 2.

© 1989, SYSTAT, Inc.

Choosing a distance metric

The [Minkowski constant] box allows you to specify the exponent (r) in a general power metric. **MDS** can use a power (Minkowski) metric in its calculations of distances between points in the configuration:

$$d_{jk} = \left[\sum_{i=1}^{p} | x_{ij} - x_{ik} |^{r} \right]^{\frac{1}{r}}$$

In this equation, p is the number of dimensions. For ordinary Euclidean distance, r=2. This is the default value.

If you specify r=1, the distance metric is called "city block distance" since distances between points in the scaling configuration are measured along "streets" parallel to the reference axes. For values other than 1 or 2 computation is slower, because logarithms and exponentials are used.

Shepard plots

Shepard diagrams plot distances in the final configuration against dissimilarities (or similarities). Normally, you get one Shepard diagram. If you specify two, you get one diagram for the first point in each distance pair and one for the second. Furthermore, the fitted regression function itself is plotted as asterisks in the scatterplot.

Choosing a loss function

Normally, **MDS** uses Kruskal's stress formula 1 (Kruskal, 1964). You can also select Guttman's coefficient of alienation (Guttman, 1968). To do this, click either the [Guttman] or [Kruskal] buttons.

Iterations with Kruskal's method are faster but usually take longer to converge to a minimum value than those with the Guttman method. The procedure used in the latter has been found in simulations to be less susceptible to local minima than that used in the Kruskal method (Lingoes and Roskam, 1973).

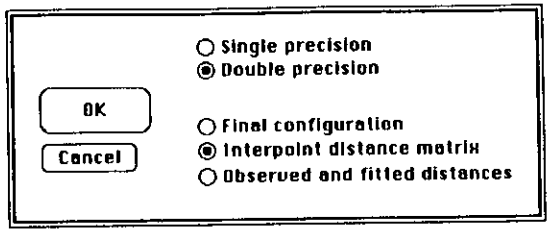
Choosing a regression function

If you use the Kruskal scaling method, you may choose to fit a monotonic or linear function of distances onto input dissimilarities. The standard option is monotonic multidimensional scaling. To avoid degenerate solutions, however, linear scaling is sometimes handy.

Saving files

You may save the matrix of distances between points in the final scaled configuration by clicking the [Save file] option. After you specify a filename, a dialog box lets you choose to save the final configuration, a matrix of distances between points, or the observed and fitted distances.

© 1989, SYSTAT, Inc.

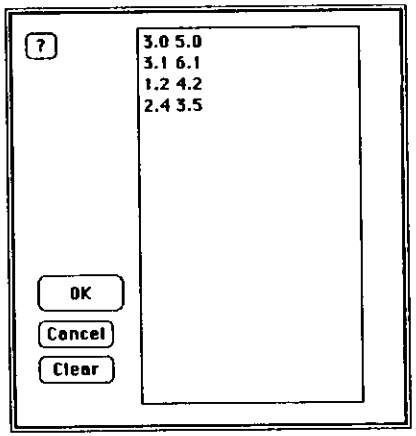


Click [Observed and fitted distances] to save the data (in a variable named DATA), distances (DIST), and estimated distances (DHAT) in a rectangular SYSTAT file. If you choose this option, **MDS** displays the root mean squared residuals for each point in its output. Since Stress is a function of the sum of squared residuals, the root mean squared residuals are a measure of the influence of each point on the Stress statistic. This can help you to identify ill-fitting points.

Configuration...

If you are having problems with degenerate solutions (don't we all in today's world?), you may want to specify your own starting configuration for the points by using **Configuration...**

- Select **MDS/Configuration...** from the **Stats** menu
- Enter a matrix of starting configurations in the large box; separate columns with spaces, and press Return between rows
- Click OK



© 1989, SYSTAT, Inc.

SYSTAT reads as many numbers in each row as you specify for [Dimensions] in the **Scale...** dialog box. It reads as many rows as there are points to scale. Check the configuration in your output to be sure you entered it correctly.

Confirmatory scaling

You can specify a configuration for confirmatory analysis. Enter a hypothesized configuration and let the program iterate only once (specify 1 for [Iterations] when you use **Scale...**). Then look at the Stress.

© 1989, SYSTAT, Inc.

Computation

Missing data

Missing values in a similarity/dissimilarity matrix are ignored in the computation of the loss function that determines how points in the configuration are moved. See the discussion of algorithms for information on how this function is computed.

If you compute a similarity matrix with **Corr** for input to **MDS**, the matrix will have no missing values unless all of your cases in the raw data have a constant or missing value on one or more variables.

Algorithms

This section summarizes algorithms separately for the Kruskal and Guttman methods. The algorithms in these options substantially follow those of Kruskal (1964ab) and Guttman (1968). **MDS** output should agree with other nonmetric multi-dimensional scaling except for rotation, dilation, and translation of the configuration. Secondary documentation can be found in Schiffman, Reynolds, and Young, (1981), and the other multidimensional scaling references. The summary assumes that dissimilarities are input. If similarities are input, **MDS** inverts them.

Kruskal method

The program begins by generating a configuration of points whose interpoint distances are a linear function of the input data. For this estimation, **MDS** uses a metric multidimensional scaling. Missing values in the input dissimilarities matrix are replaced by mean values for the whole matrix. Then the values are converted to distances by adding a constant. A scalar products matrix **B** is then calculated following the procedures described in Torgerson (1958). The initial configuration matrix **X** in p dimensions is computed from the first p eigenvectors of **B** using the Young-Householder procedure (Torgerson, 1958)

Once an initial configuration is computed by the metric method, nonmetric optimization begins (there are no metric pre-iterations). At the beginning of each iteration, the configuration is normalized to have zero centroid and unit dispersion. Next, Kruskal's DHAT (fitted) distance values are computed by a monotonic regression of distances onto data. Tied data values are ordered according to their corresponding distances in the configuration.

© 1989, SYSTAT, Inc.

Stress (formula 1) is calculated from fitted distances, observed distances, and input data values. If the stress is less than .001, or has decreased in the last 5 iterations less than .001 per iteration, or the number of iterations equals the number specified by the user (default is 50), iterations terminate (i.e., go to the next paragraph). Otherwise, the negative gradient is computed for each point in the configuration by taking the partial derivatives of stress with respect to each dimension. Points in the configuration are moved along their gradients with a step size chosen as a function of the rate of descent: the steeper the descent, the smaller the step size. This completes an iteration.

After the last iteration, the configuration is shifted so that the origin lies in the centroid. Thus, the point coordinates sum to zero on each dimension. Moreover, the configuration is normalized to unit size so that the sum of squares of its coordinates is 1. If the [Minkowski constant] is 2 (Euclidean scaling, which is the standard option), the final configuration is rotated to its principal axis.

Guttman method

The initial configuration for the Guttman option is computed according to Lingoes and Roskam (1973). Principal components are computed on a matrix C, where

$$c_{ij} = 1 - \frac{r_{ij}}{\frac{n(n-1)}{2}}$$

r_{ij} are the ranks of the input dissimilarities (smallest rank corresponding to smallest dissimilarity), and n is the number of points. The diagonal elements of C are

$$c_{ij} = 1 - \sum r_{ij}$$

where the sum is taken over the entire row of the dissimilarity matrix.

© 1989, SYSTAT, Inc.

For the iteration stage, the initial configuration is normalized as in the Kruskal method. Then rank images corresponding to each distance in the configuration are computed by permuting the configuration distances so that they mirror the rank order of the original input dissimilarities. Ties in the data are handled as in the Kruskal method. These rank images are used to compute the Guttman/Lingoes coefficient of alienation. Iterations are terminated if this coefficient becomes arbitrarily small, if the number of iterations exceeds the maximum, or if the change in its value becomes small. Otherwise, the points in the configuration are moved five times using the same rank images but different interpoint distances each time to compute a new negative gradient. These five cycles within each iteration are what lengthens the calculations in the Guttman method. This completes an iteration.

The final configuration is rotated and scaled as with the Kruskal method. Guttman/Lingoes programs normalize the extreme values of the configuration to unity and thus do not plot the configuration with a zero centroid, so **MDS** output corresponds to their output within rigid motion and configuration size.

© 1989, SYSTAT, Inc.

5.1 Guttman method

The following data from Ekman (1954) are judged similarities among 14 different spectral colors. The judgments were averaged across 31 subjects. The variable labels cover the wavelengths in nanometers of the light samples. Reds are in the 400's and blue-violets in the 600's. The larger the numbers, the more similar the colors were judged. The file (EKMAN) has no diagonal elements and has type Similarity.

W434	W445	W465	W472	W490	W504	W537	W555	W584	W600	W610	W628	W651	W674
.													
86	.												
42	50	.											
42	44	81	.										
18	22	47	54	.									
06	09	17	25	61	.								
07	07	10	10	31	62	.							
04	07	08	09	26	45	73	.						
02	02	02	02	07	14	22	33	.					
07	04	01	01	02	08	14	19	58	.				
09	07	02	00	02	02	05	04	37	74	.			
12	11	01	01	01	02	02	03	27	50	76	.		
13	13	05	02	02	02	02	02	20	41	62	85	.	
16	14	03	04	00	01	00	02	23	28	55	68	76	.

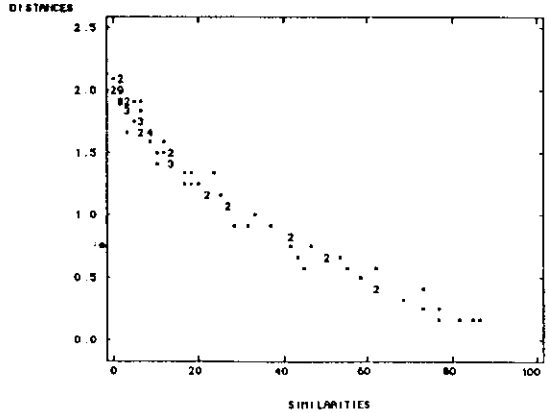
We will scale these judgments in two dimensions to see if the data fit a perceptual "color wheel."

- Select **MDS/Scale...** from the **Stats** menu
- Choose [Guttman]
- Click OK

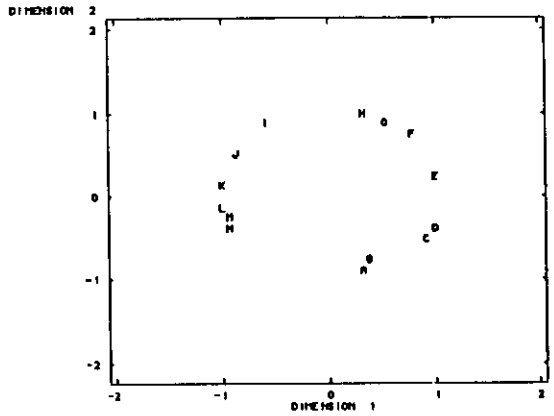
© 1989, SYSTAT, Inc.


```
Monotonic Multidimensional Scaling

Minimizing Guttman/Lingoes Coefficient of Alienation in 2 Dimensions

Iteration  Alienation
---------  ----------
        0       .060
        1       .047
        2       .044
        3       .043
        4       .043
        5       .043
        6       .043
        7       .043
        8       .043
        9       .043
       10       .043
       11       .043
       12       .043

Stress of Final Configuration is:       .04259

 Coordinates in 2 Dimensions

Variable    Plot   Dimension
--------    ----   ---------
                       1      2
    W434      A      .31   -.89
    W445      B      .40   -.84
    W465      C      .91   -.55
    W472      D      .97   -.49
    W490      E      .98    .13
    W504      F      .79    .63
    W537      G      .53    .86
    W555      H      .34   1.00
    W584      I     -.58    .76
    W600      J     -.84    .40
    W610      K     -.98    .03
    W628      L     -.99   -.21
    W651      M     -.94   -.35
    W674      N     -.91   -.48
```

SHEPARD DIAGRAM



© 1989, SYSTAT, Inc.



The fit of configuration distances to original data is extremely close, as evidenced by the low coefficient of alienation and clean Shepard diagram. The resulting configuration is almost circular, denoting a "circumplex" by Guttman (1954). There is a large gap at the bottom of the figure, however, because the perceptual color between deep red and dark purple is not a spectral color.

© 1989, SYSTAT, Inc.

5.2 Kruskal method

The following data, adapted from an experiment by Rothkopf (1957), were originally obtained from 598 subjects who judged whether or not pairs of Morse code signals presented in succession were the same. Morse code signals for letters and digits were used in the experiment, and all pairs were tested in each of two possible sequences. For multidimensional scaling, the data for letter signals have been averaged across sequence and the diagonal (pairs of the same signal) has been omitted. The data in this form were first scaled by Shepard (1963).

```
A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y  Z

.
4
5 37  .
10 46 17  .
4  9  9  6  .
9 39 31 21  2
9 17 20 39  2 12  .
8 33 15 30  6 30  7  .
55  6  9 11 13  6  5 11  .
6 14 28 11  1 28 20  7  3  .
13 29 31 77  3 16 29 10  5 24  .
2 54 39 50  8 37 14 27 12 30 30  .
24  9  9 11  5 12 21  8 10  9 16  7  .
32  7  8 18  6  9 11 12 21  3 14  5 60  .
6  9 14  7  5  9 69  7  4 30 28  9  7  6  .
5 22 42 19  3 39 37 13  3 70 14 41  7  7 29  .
8 22 36 10  4 13 16  6  3 37 27 31 11  5 28 47  .
24 15 20 34  7 34 23 16 13 11 21 30 17 20 11 26  9  .
20 21  9 29  9 20  5 48 25  2  8 11  6 14  4  7  2 17  .
9  6  3  5 56  2  4  5 15  6  8  6  7  7  5  7  3  3  8  .
25 23  9 24  3 26 13 35 15  8 33 24 12 14  8 16  9 31 53  7  .
9 25 17 18  6 28 10 49  5 10 21 31  4  2 11 18 12 20 17  5 44  .
13 14 22 24  5 30 28 12  6 17 31 19 18  9 25 19 18 61 12  4 26 18  .
9 74 38 29  4 23 16 19  1 28 56 36  9  6 29 25 49 12 11  2 14 37 21  .
8 26 72 15  5 26 22 11  3 31 14 19 11  6 20 35 51  8  6  3  7 24 23 46  .
3 44 41 25  3 17 28 10  7 25 19 55  9  6 21 44 67 13  5  4 10 10 17 31 32  .
```

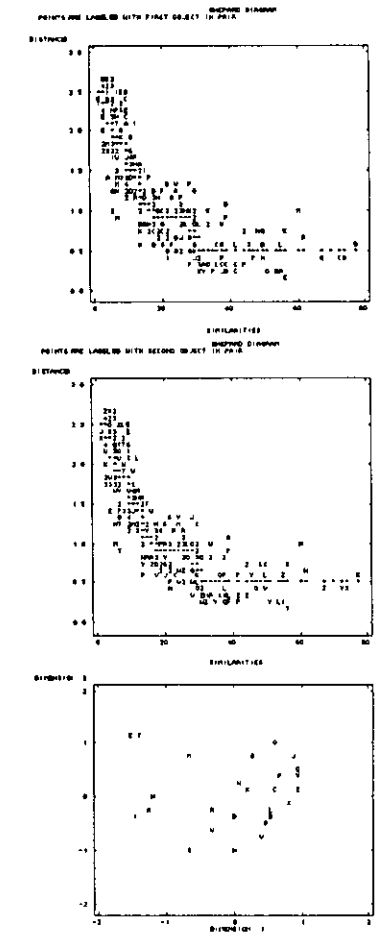
- Select **MDS/Scale...** from the **Stats** menu
- Specify 2 for [Shepard plots]
- Click OK

Here is the first part of the output. The solution required 17 iterations. The stress statistic typically declines smoothly across iterations. If it fluctuates across iterations, you may want to try another starting configuration.

© 1989, SYSTAT, Inc.

```
Monotonic Multidimensional Scaling

Minimizing Kruskal Stress (FORM 1) in 2 Dimensions

Iteration  Stress
---------  ------
    0       .260
    1       .235
    2       .216
    3       .200
    4       .189
    5       .183
    6       .181
    7       .179
    8       .178
    9       .177
   10       .177
   11       .176
   12       .176
   13       .176
   14       .176
   15       .176
   16       .176
   17       .175

Stress of Final Configuration is:      .17541

 Coordinates in 2 Dimensions

Variable    Plot   Dimension
--------    ----   ---------
                      1      2
       A      A    -1.24   -.36
       B      B      .50   -.41
       C      C      .63    .03
       D      D      .02   -.38
       E      E    -1.52   1.02
       F      F      .45   -.57
       G      G      .27    .66
       H      H      .01  -1.02
       I      1    -1.46   -.39
       J      J      .85    .69
       K      K      .19    .09
       L      L      .55   -.25
       M      M     -.64    .66
       N      N    -1.19   -.11
       O      O      .60    .99
       P      P      .67    .26
       Q      Q      .93    .48
       R      R     -.30   -.29
       S      S     -.67  -1.02
       T      T    -1.43   1.04
       U      U     -.33   -.70
       V      V      .38   -.84
       W      W      .05    .22
       X      X      .82   -.19
       Y      Y      .91    .31
       Z      Z      .94    .10
```


© 1989, SYSTAT, Inc.



© 1989, SYSTAT, Inc.

SYSTAT plots the Shepard diagrams first. These are plots of distances against input dissimilarities or similarities. If we choose two Shepard diagrams and there are no missing data and not too many points, the computed regression function is marked by asterisks and two Shepard diagrams are printed, as in this case. In monotonic scaling, the regression function has steps at various points. For most solutions, the function in this plot should be relatively smooth (without large steps). If the function looks like one or two large steps, you should consider choosing linear scaling for [Regression].

Two Shepard diagrams are printed so we can identify which pairs of points in the final configuration have poor fits to the original data. In the first plot, each distance is identified by the plotting character of the first member of each pair. In the second plot, the same distances are identified by the second member. In this example, there are many overlapping points.

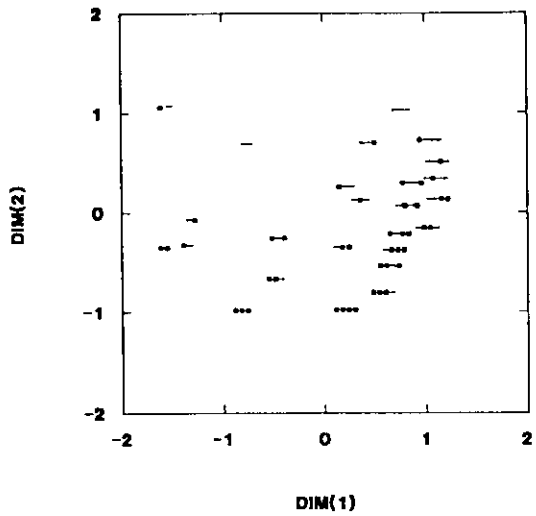
Notice that E frequently appears at the edge of the scatterplots. This indicates that the similarities (confusions) of the Morse Code E with the other symbols are fit relatively poorly by the two dimensional scaling solution. E would likely (though not necessarily) have a substantial third coordinate in a three dimensional solution.

SYSTAT prints the coordinates of the solution and then plots them for every pair of dimensions. Plotting characters are letters of the alphabet (in this case, the same as the variable names). You can see this plot at the bottom of the 3 panels.

The solution has been rotated to principal axes (i.e. the major variation is on the first dimension). This rotation is not performed unless the scaling is in Euclidean space, as in the present example.

The following plot is not provided by **MDS**. We replaced the letters in the previous plot with their corresponding Morse code symbols in a character variable and used SYGRAPH to plot the configuration in high resolution. The two dimensional solution clearly distinguishes short signals from long and dots from dashes.

© 1989, SYSTAT, Inc.



Notice that dashes tend to appear in the upper right and dots in the lower left. Long codes tend to appear in the lower right and short in the upper left.

© 1989, SYSTAT, Inc.

6 MGLH

Introduction 118
Simple linear models 118
Regression 120
Least squares 122
Estimation and inference 122
Standard errors 124
Hypothesis testing 124
The multiple correlation 126
Regression diagnostics 126
Multiple regression 130
Variable selection 132
Multivariate regression 134
Matrix notation 134
The multivariate general linear hypothesis 135
Analysis of variance 136
Two-group ANOVA 136
Multi-group ANOVA 138
Multivariate ANOVA 139
Conclusion 139
Computation 140
Missing data 140
Algorithms 140
Regression 143
Introduction 144
Usage 145
Examples 158
Analysis of variance 180
Introduction 181
Usage 182
Examples
Multivariate models 282
Introduction 283
Usage 284
Examples 292

Detailed outlines for the "Regression," "Analysis of variance," and "Multivariate models" sections appear at the beginning of each section.

© 1989, SYSTAT, Inc.

MGLH 6

Overview

MGLH stands for **M**ultivariate **G**eneral **L**inear **H**ypothesis. It can estimate and test any univariate or multivariate general linear model, including those for multiple regression, analysis of variance or covariance, and other multivariate procedures such as discriminant analysis and principal components. All these procedures are minor variants of a single mathematical model. While it is not necessary to learn all the details of this model in order to use these procedures, you *can* learn how they are related so that you will be able to use any of these procedures without learning new terminology for each.

Following the "Introduction," which shows how these procedures are related, the remainder of the chapter is divided into three sections: "Regression," "Analysis of variance," and "Multivariate models." The principal methods discussed within these sections are:

Regression	***Analysis of Variance***	***Multivariate Models***
simple linear	one-way	discriminant analysis
multiple linear	two-way	principal components
stepwise	factorial	canonical correlation
polynomial	nested	
mixture	unbalanced	
weighted	post-hoc tests	
	mixed	
	repeated measures	
	split plot	
	multivariate	

© 1989, SYSTAT, Inc.

Linear models are models based on *lines*. More generally, they are based on linear surfaces, such as lines, planes, and hyperplanes. Linear models are widely applied because lines and planes often appear to describe well the relations among variables measured in the real world. We will begin by examining the equation for a straight line, and then move to more complex linear models. Finally, we will see that the procedure called Analysis of Variance is simply a linear model applied to coded categorical data.

Simple linear models

A linear model looks like this:

$$y = a + bx$$

This is the equation for a straight line you learned in school. The quantities in this equation are:

- y a dependent variable
- x an independent variable

Variables are quantities that can vary (have different numerical values) in the same equation. The remaining quantities are called *parameters*. A parameter is a quantity that is constant in a particular equation, but that can be varied to produce other equations in the same general family.

Here are the parameters:

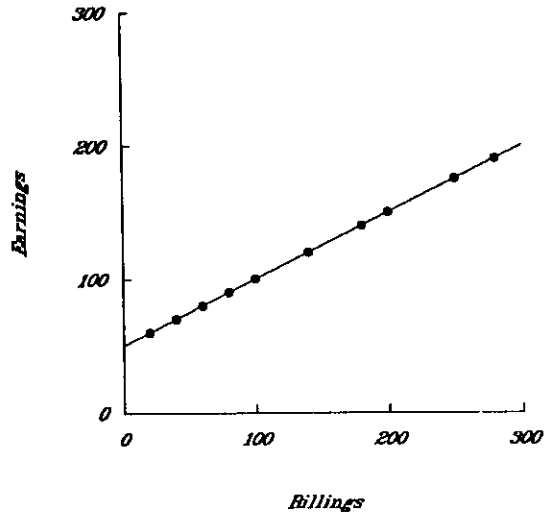
- a the value of y when x is zero. This is sometimes called a *y-intercept* (where a line intersects the y axis in a graph when x is zero).
- b the slope of the line, or the number of units y changes when x changes by one unit.

Let's look at an example now to see how this works. Here are some data showing the yearly amount a partner should make in a certain small accounting firm based on annual billings (both in thousands of dollars).

© 1989, SYSTAT, Inc.

EARNINGS	BILLINGS
60	20
70	40
80	60
90	80
100	100
120	140
140	180
150	200
175	250
190	280

We can plot these data with EARNINGS on the vertical axis (dependent variable) and BILLINGS on the horizontal (independent variable). Notice in the following figure that all the points lie on a straight line



© 1989, SYSTAT, Inc.

What is the equation for this line? Look at the vertical axis value on the sloped line where the independent variable has a value of zero. Its value is 50. An accountant is paid fifty thousand dollars even when billing nothing. Thus, *a* is 50 in our equation. What is *b*? Notice that the line rises by ten thousand dollars when billings change by twenty thousand. The line rises half as fast as it runs. You can also look at the data and see that the earnings change by 1 dollar as billing changes by 2 dollars. Thus, *b* is .5, or a half, in our equation.

Why bother with all these calculations? We could use the table to determine an accountant's compensation, but the formula and the line graph allow us to determine wages *not* found in the table For example, we now know that $30,000 in billings would yield earnings of $65,000:

$$\text{EARNINGS} = 50000 + .5 \times 30000 = 65000$$

When we do this, however, we must be sure that we can use the same equation on these new values We must be careful when *interpolating*, or estimating, wages for billings between the ones we have been given. Does it make sense to compute earnings for $25,000 in billings, for example? It probably does. Similarly, we must be careful when *extrapolating*, or estimating from units outside of the range of values we have been given. What about negative billings, for example? Would we want to pay an embezzler? Be careful. Equations and graphs usually are meaningful only within or close to the *range* of *y* values and *domain* of *x* values in the data.

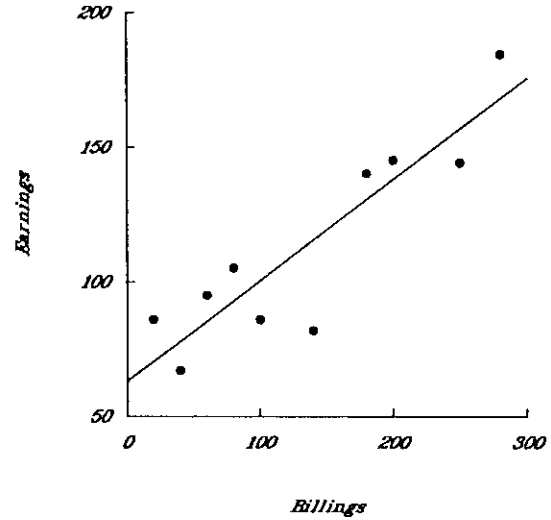
Regression

Now, data are seldom this clean unless we design them to be that way. Accounting firms typically fine tune their partners' earnings according to many factors. Here are the actual billings and earnings for our small firm:

EARNINGS	BILLINGS
86	20
67	40
95	60
105	80
86	100
82	140
140	180
145	200
144	250
184	280

© 1989, SYSTAT, Inc.

Our techniques for computing a linear equation won't work with these data. Look at the following graph. There is no way to draw a straight line through all the data.



Given the irregularities in our data, the line drawn in the figure is a compromise. How do we find a best fitting line? If we are interested in predicting earnings from the billing data values rather well, a reasonable method would be to place a line through the points such that the vertical deviations between the points and the line (errors in predicting earnings) are as small as possible In other words, these deviations (absolute discrepancies, or *residuals*) should be small, on average, for a good fitting line.

© 1989, SYSTAT, Inc.

The procedure of fitting a line or curve to data such that residuals on the dependent variable are minimized in some way is called *regression*. Because we are minimizing vertical deviations, the regression line often appears to be more horizontal than we might place it by eye, especially when the points are fairly scattered. It "regresses" toward the mean value of *y* across all the values of *x*, namely, a horizontal line through the middle of all the points. Remember, the regression line is not intended simply to pass through as many points as possible. It is for predicting the dependent variable as accurately as possible given each value of the independent variable.

Least squares

There are several ways to draw the line such that, on average, the deviations are small. We could minimize the *mean*, the *median*, or some other measure of the typical behavior of the absolute residuals. Or, we can minimize the sum (or mean) of the squared residuals, which yields almost the same line in most cases. Using squared instead of absolute residuals gives more influence to points whose *y* value is farther from the average of all *y* values, which is not always desirable, but it makes the mathematics a lot simpler. This method, used by **MGLH** and almost every linear modeling program, is called *ordinary least squares*.

By selecting EARNINGS as the dependent variable and BILLINGS as the independent variable in an **MGLH** dialog box or MODEL statement (if you use the optional command interface), we can compute the ordinary least squares regression *y*-intercept as 62.8 thousand dollars and the slope as .375. These values do not predict any single accountant's earnings exactly. They describe the whole firm well in the sense that, on the average, the line predicts a given earnings value fairly closely from a given billings value.

Estimation and inference

We often want to do more with such data than draw a line on a picture. In order to generalize, formulate a policy, or test an *hypothesis*, we need to make an *inference*. Making an inference implies that we think a model describes a more general *population* from which our data have been randomly *sampled*. In the present example, this population is all possible accountants who might work for this firm. Sometimes it is difficult to imagine a population from which you have gathered data. A population can be "all possible voters," "all possible replications of this experiment," or "all possible moviegoers." When you make inferences, you should have a population in mind.

© 1989, SYSTAT, Inc.

To make an *inference* about compensation we need to construct a linear model for our population that includes a parameter for random error. In addition, we need to change our notation to avoid confusion later. We are going to use Greek to denote parameters and italic Roman letters for variables. The error parameter is usually called ε:

$$y = \alpha + \beta x + \varepsilon$$

Notice that ε is a *random variable*. It varies like any other variable (e.g. *x*), but it varies randomly, like the tossing of a coin. Since ε is random, our model forces *y* to be random as well, because adding various values (α and βx) to a random variable (ε) produces another random variable. In ordinary language, we are saying with our model that earnings are only partly predictable from billings. They vary slightly according to many other factors, which we assume are random.

We don't know all the factors governing the firm's compensation decisions, but we assume:

1) all the salaries are derived from the same linear model,
2) the error in predicting a particular salary from billings using the model is *independent* of (not predictable from) the error in predicting other salaries, and
3) the errors in predicting all the salaries come from the same random distribution.

Our model for predicting in our population contains parameters, but unlike our perfect straight line example, we cannot compute these parameters directly from the data. The data we have are only a small *sample* from a much larger population, so we can only *estimate* the parameter values using some statistical method on our sample data. Those of you who have heard this story before may not be surprised that ordinary least squares is one reasonable method for estimating parameters when our three assumptions are appropriate. Without going into all the details, we can safely assume that if our population assumptions are true and if we *randomly sample* some cases (i.e. each sample of fixed size has an equal chance of being chosen) from the population, the least squares estimates of α and β will, on average, be close to the values of α and β in the population.

© 1989, SYSTAT, Inc.

So far, we have done what seems like a sleight of hand. We delved into some abstruse language and came up with least squares values for *a* and *b* just as before. There is something new, however. We have now added conditions that define our least squares values as sample estimates of population values. We now regard our sample data as one instance of many possible samples. Our compensation model is like Plato's cave metaphor: we think it typifies how this small firm makes compensation decisions about *any* accountant, not just the ones we sampled. Before, we were computing *descriptive statistics* about a sample. Now we are computing *inferential statistics* about a population.

Standard errors

There are several inferential statistics relevant to the estimation of α and β we have just done. Perhaps most important is a measure of how variable we could expect our estimates to be if we continued to sample data from our population and use lease squares. A statistic calculated by **MGLH** shows what we could expect this variation to be. It is called, appropriately, the *standard error of estimate*, or "Std Error" in the output. The standard error of the y-intercept, or *regression constant* is in the first row of the coefficients: 10.440. The standard error of the billing coefficient is 0.065. Look for these numbers in the following output:

```
Dep Var:EARNINGS      N:   10    Multiple R:  .897    Squared Multiple R:  .804
Adjusted squared multiple R:  .779     Standard error of estimate:       17.626

  Variable    Coefficient    Std Error     Std Coef Tolerance    T    P(2 tail)

CONSTANT          62.838       10.440          0.000 .          6.019    0.000
 BILLINGS          0.375        0.065          0.897 1.0000000  5.728    0.000

                          Analysis of Variance

   Source    Sum-of-Squares    DF  Mean-Square     F-Ratio       P

Regression      10191.109       1    10191.109       32.805      0.000
  Residual       2485.291       8      310.661
```

Hypothesis testing

From these standard errors, we can construct *hypothesis tests* on these coefficients. Suppose a skeptic approached us and said, "Your estimates look as if something is going on here, but actually this firm computes salaries by throwing dice. Nobody cares about billings. You just happened to pick a sample that gives the impression that billings matter. It was the luck of the draw that provided you with such a misleading picture. In reality, both α and β are zero in the population because billings play no role in determining earnings." We can reply: *if* it were luck of the draw and *if* the firm really does toss dice when doing compensation, would it be likely for us to find a coefficient estimate in a sample at least this different from zero?

© 1989, SYSTAT, Inc.

To represent these alternatives as a bet between us and the skeptic, we must agree on some *critical level* for deciding who will win the bet. If the likelihood of a sample result at least this extreme occurring by chance is less than or equal to this critical level (say, five times out of a hundred), we win. Otherwise, the skeptic wins.

This logic might seem odd at first since in almost every case, our skeptic's *null hypothesis* would appear ridiculous, and our *alternative hypothesis* (that the skeptic is wrong) seems plausible. Two scenarios are relevant here, however. The first is the lawyer's. We are trying to make a case here. The only way we will prevail is if we convince our skeptical jury beyond a reasonable doubt. In statistical practice, that reasonable doubt level is relatively liberal: fewer than five times in a hundred. The second scenario is the scientist's. We are going to stake our reputation on our model. If someone sampled more data and failed to find nonzero coefficients, much less coefficients similar to ours, no one would pay attention to us in the future.

To compute probabilities, we must count all possibilities or refer to a mathematical probability distribution that approximates these possibilities well. The most widely used approximation is the *normal curve*, which we reviewed briefly in the introduction to this manual. For large samples, the regression coefficients will tend to be normally distributed under the assumptions we made above. To allow for smaller samples, however, we will add to our list of assumptions the condition that

4) the errors in predicting the salaries come from a *normal distribution*.

If we estimate the standard errors of the regression coefficients from the data instead of knowing them in advance, then we should use the *t* distribution instead of the normal. The two-tail value for the probability represents the area under the theoretical *t* probability curve corresponding to coefficient estimates whose absolute values are more extreme from zero than the ones we obtained. For both parameters in the model of accountants' earnings, these values (given as P(2 tail)) are less than .001, leading us to reject our null hypothesis at well below the .05 level.

© 1989, SYSTAT, Inc.

Below this part of our output, we get an analysis of variance table that tests the significance of our entire model. The null hypothesis corresponding to the F ratio (32.805) and its associated p value is that the billing variable coefficient is equal to zero. This test overwhelmingly rejects the null hypothesis ($p<.001$).

The multiple correlation

In the same output is a statistic called the ***squared multiple correlation.*** This is the proportion of the total variation in the dependent variable (earnings) accounted for by the linear prediction using billings. The value here (.804) tells us that approximately 80 percent of the variation in earnings can be accounted for by a linear prediction from billings. The rest of the variation, as far as this model is concerned, is random error. The square root of this statistic is called, not surprisingly, the multiple correlation The adjusted squared multiple correlation (.779) is what we would expect the squared multiple correlation to be if we used the model we just estimated on a new sample of ten accountants in the firm. It is smaller than the squared multiple correlation because the coefficients were optimized for this sample rather than for the new one.

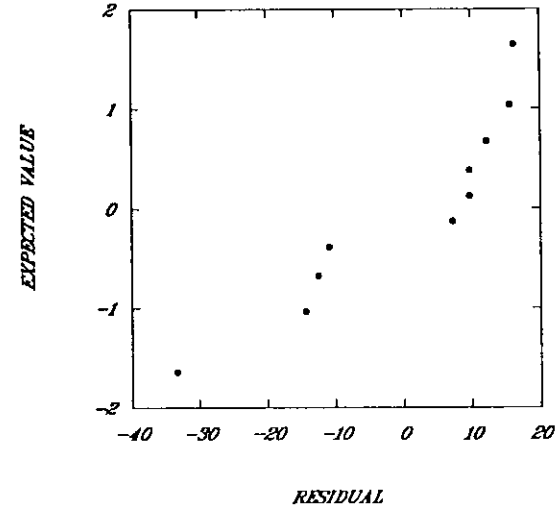
Regression diagnostics

We do not need to understand the mathematics of how a line is fitted by ordinary least squares to use regression. You can fit a line to any x-y data by the method of least squares. The computer doesn't care where the numbers come from. To have a model and estimates that mean something, however, you should be sure the assumptions are reasonable and that the sample data appear to be sampled from a population that meets the assumptions.

The sample analogues of the errors in the population model are the *residuals*—the differences between the observed and predicted values of the dependent variable. There are many diagnostics you can perform on the residuals. Here are the most important ones:

1) **The errors are normally distributed.** Draw a normal probability plot (PPLOT) of the residuals:

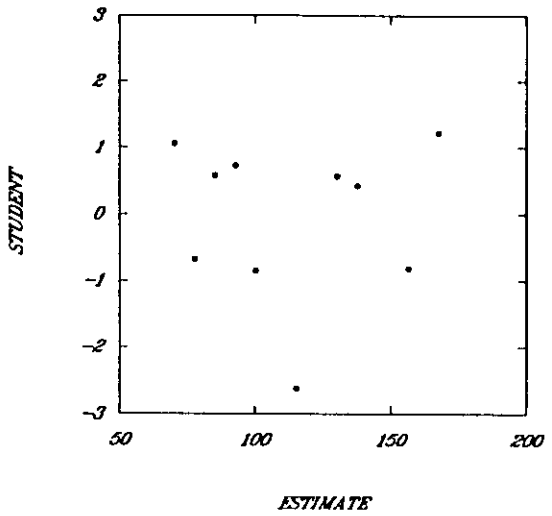
© 1989, SYSTAT, Inc.



The residuals should fall approximately on a diagonal straight line in this plot. When the sample size is small, as in our accounting example, the line may be quite jagged. It is difficult to tell by any method whether a small sample is from a normal population. You can also plot a histogram or stem-and-leaf diagram of the residuals to see if they are lumpy in the middle with thin, symmetric tails.

2) **The errors have constant variance.** Plot the residuals against the estimated values. The following plot shows *studentized residuals* (STUDENT) against estimated values (ESTIMATE). Studentized residuals follow approximately a t distribution, which means we can spot large values (greater then 2 or 3 in absolute magnitude) as possible problems.

© 1989, SYSTAT, Inc.



Our residuals should be arranged in a horizontal band within 2 or 3 units around zero in this plot. Again, since there are so few observations, it is difficult to tell whether they violate this assumption in this case. There is only one particularly large residual, and it is toward the middle of the values. This accountant billed 140 thousand and is earning only 80. He or she might have a gripe about supporting a higher share of the firm's overhead.

3) **The errors are independent.** Several plots can be done. Look at the plot of residuals against estimated values above. Make sure that the residuals are randomly scattered above and below the zero horizontal and that they do not track in a snaky way across the plot. If they look as if they were shot at the plot by a horizontally moving machine gun, then they are probably not independent of each other. You may also want to plot residuals against other variables, such as time, orientation, or other ways that might influence the variability of your dependent measure. **ACF plot...** in **Series** measures whether the residuals are *serially correlated*. Here is an autocorrelation plot:

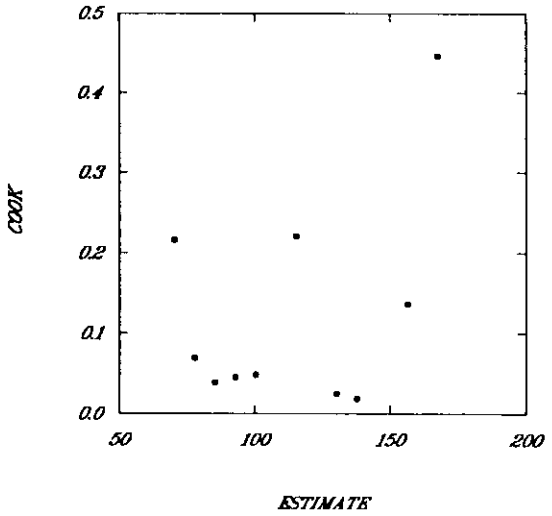
© 1989, SYSTAT, Inc.

Plot of Autocorrelations

Lag	Corr	SE	-1.0 -.8 -.6 -.4 -.2 .0 .2 .4 .6 .8 1.0
1	-.162	.316	(**** |)
2	-.366	.324	(********* |)
3	.246	.363	(| ******)
4	.018	.380	(|)
5	-.379	.380	(********* |)
6	.061	.416	(| *)
7	.164	.417	(| ****)
8	.150	.423	(| ***)
9	.103	.428	(| **)
10	.000	.431	(|)

All the horizontal bars should be within the parentheses if each residual is not predictable from the one preceding it, and the one preceding that, and the one preceding that, and so on.

4) **All the members of the population are described by the same linear model.** Plot Cook's distance (COOK) against the estimated values:



© 1989, SYSTAT, Inc.

Cook's distance measures the influence of each sample observation on the coefficient estimates. Observations that are far from the average of all the independent variable values or that have large residuals tend to have a large Cook's distance value (say, greater than 2). Cook's D actually follows closely an *F* distribution, so aberrant values depend on the sample size. You can read more about it later in this chapter. We don't want to find a large Cook's D value for an observation because it would mean that the coefficient estimates would change substantially if we deleted that observation. While none of the Cook values are extremely large in our example, could it be that the largest one in the upper right corner is the founding partner in the firm? Despite large billings, this partner is earning more than the model predicts.

In conclusion, keep in mind that all our diagnostic tests are themselves a form of inference. We can assess theoretical errors only through the dark mirror of our observed residuals. Despite this caveat, testing assumptions graphically is critically important. You should never publish regression results until you have examined these plots.

Multiple regression

A multiple linear model has more than one independent variable:

$$y = a + bx + cz$$

This is the equation for a plane in three dimensional space The parameter a is still an intercept term. It is the value of y when x and z are zero. The parameters b and c are still slopes. One gives the slope of the plane along the x dimension and the other along the z dimension.

The statistical model has the same form:

$$y = \alpha + \beta x + \gamma z + e$$

Before we run out of letters for independent variables, let's switch to a more frequently used notation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

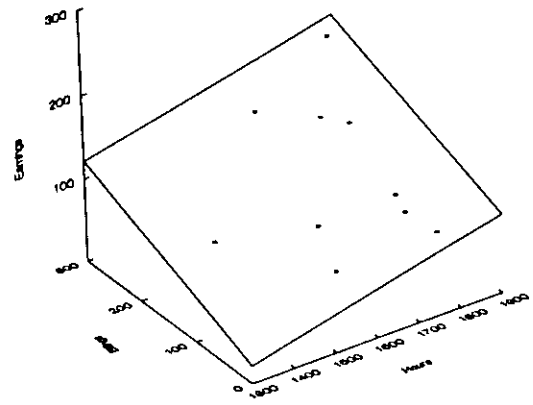
Notice that we are still using Greek letters for parameters and Roman letters for variables.

© 1989, SYSTAT, Inc.

Now, let's look at our accounting data again. We have learned that there is another variable that appears to determine earnings—the number of hours billed per year by each accountant. Here is an expanded listing of the data:

EARNINGS	BILLINGS	HOURS
86	20	1771
67	40	1556
95	60	1749
105	80	1754
86	100	1594
82	140	1400
140	180	1780
145	200	1737
144	250	1645
184	280	1863

For our model, β_1 is the coefficient for BILLINGS and β_2 is the coefficient for HOURS. Let's look first at its graphical representation. The following figure shows the plane fit by least squares to the points representing each accountant. Notice how the plane slopes upward on both variables. BILLINGS and HOURS both contribute positively to EARNINGS in our sample.



© 1989, SYSTAT, Inc.

Fitting this model involves no more work than fitting the simple regression model. We specify one dependent and two independent variables for **MGLH** and estimate the model as before. Here is the result:

```
Dep Var:EARNINGS      N:   10    MULTIPLE R:  .998    Squared Multiple R:  .996
Adjusted squared Multiple R:  .995     Standard Error of Estimate:         2.678

  Variable     Coefficient     Std Error     Std Coef Tolerance     T     P(2 tail)

CONSTANT          -139.925        11.116         0.000  .          -12.588    0.000
BILLINGS             0.333         0.010         0.797  .9510698   32.690    0.000
   HOURS             0.124         0.007         0.449  .9510698   18.429    0.000

                              Analysis of Variance

   Source    Sum of Squares    DF  Mean Square    F ratio       P

Regression       12626.210     2     6313.105     880.493     0.000
  Residual          50.190     7        7.170
```

This time, we have one more row in our regression table for HOURS. Notice that its coefficient (.124) is smaller than that for BILLINGS (.333). This is due partly to the different scales of the variables. HOURS are measured in larger numbers than BILLINGS. If we wish to compare the influence of each independent of scales, we should look at the standardized coefficients. Here, we still see that BILLINGS (.797) play a greater role in predicting EARNINGS than do HOURS (.449). Notice also that both coefficients are highly significant and that our overall model is highly significant, as shown in the analysis of variance table.

Variable selection

In applications, you may not know which subset of predictor variables in a larger set constitute a "good" model. Strategies for identifying a good subset are many and varied: forward selection, backward elimination, stepwise (either a forward or backward type), and all subsets. Forward selection begins with the "best" predictor, adds the next "best" and continues entering variables to improve the fit. Backward selection begins with all candidate predictors in an equation and removes the least useful one at a time as long as the fit is not substantially "worsened." Stepwise begins as either forward or backward, but allows "poor" predictors to be removed from the candidate model or "good" predictors to re-enter the model at any step. Finally, all subsets methods compute all possible subsets of predictors for each model of a given size (number of predictors) and choose the "best" one.

© 1989, SYSTAT, Inc.

Bias and variance tradeoff

Submodel selection is a tradeoff between bias and variance. By decreasing the number of parameters in the model, its predictive capability is enhanced. This is because the variance of the parameter estimates decreases. On the other side, bias may increase because the "true model" may have a higher dimension. So we'd like to balance smaller variance against increased bias.

There are two aspects to variable selection: selecting the dimensionality of the submodel (how many variables to include) and evaluating the model selected. After you determine the dimension, there may be several alternative subsets that perform equally as well. Then knowledge of the subject matter, how accurately individual variables are measured, and what a variable "communicates" may guide selection of the model to report.

A strategy

If you are in an exploratory phase of research, you might try this version of backwards stepping. First, fit a model using all candidate predictors. Then identify the least "useful" variable, remove it from the model list, and fit a smaller model. Then evaluate your results and select another variable to remove. Continue removing variables. For a given size model, you may want to remove alternative variables: that is, first remove variable A, evaluate results, replace A and remove B, etc.

Removal criteria

Decisions about which variable to remove should be based on statistics and diagnostics in the output, especially graphical displays of these values, and your knowledge of the problem at hand. As a rough guideline, consider models that include only variables that have absolute "*t*" values well above 2.0 and "tolerance" values greater than .1. We use quotation marks here because *t* and other statistics do not have their usual distributions when you are selecting subset models.

© 1989, SYSTAT, Inc.

Evaluation criteria

There is no one test to identify the dimensionality of the best submodel. Recent research by Leo Breiman emphasizes the usefulness of cross-validation techniques involving 80 percent random subsamples. Sample 80 percent of your file, fit a model, use the resulting coefficients on the remaining 20 percent to obtain predicted values, and compute R squared for this smaller sample. In over-fitting situations, the discrepancy between the R squared for the 80 percent sample and the 20 percent sample can be dramatic.

A warning

If you do not have extensive knowledge of your variables and expect this strategy to help you to find a "true" model, you can get into a lot of trouble. Automatic stepwise regression programs cannot do your work for you. You must be able to examine graphics and make intelligent choices based on theory and prior knowledge. Otherwise, you will be arriving at nonsense. See the stepwise regression example in the "Regression Examples" section of this chapter for more information.

Multivariate regression

We may extend the multiple regression equation to several dependent variables:

$$y1 = \beta_{01} + \beta_{11}x_{11} + \beta_{21}x_{21} + \varepsilon_1$$
$$y2 = \beta_{02} + \beta_{12}x_{12} + \beta_{22}x_{22} + \varepsilon_2$$
$$y3 = \beta_{03} + \beta_{13}x_{13} + \beta_{23}x_{23} + \varepsilon_3$$
$$\vdots$$

Each equation is for predicting a different dependent variable. For example, we might want to predict earnings, office size, and expense accounts from billings and hours in our accounting firm. We can't show you a picture of the resulting predictions because they would have too many dimensions to represent on paper. Nevertheless, you should know that multivariate regression is a simple extension of multiple regression.

Matrix notation

The multivariate equations get messy when we have a lot of *x* and *y* variables. There is a simpler ***matrix*** notation mathematicians use for representing arrays of variables. You don't need to understand matrix algebra to benefit from looking at the matrix representation of multiple regression. Here it is:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

© 1989, SYSTAT, Inc.

Notice the parallels to the multivariate equations. We have collected all the *y* variables into the columns of matrix **Y**, all the *x* variables into the columns of **X**, and so on.

The multivariate general linear hypothesis

Let's now see how we test hypotheses using the ***multivariate general linear model.*** Since the matrix notation works for single variables as well as many, we can use a simple matrix equation to describe every model from simple through multivariate regression. Every hypothesis can be specified as another type of simple matrix equation:

$$\mathbf{ABC'} = \mathbf{D}$$

In addition to the **B** matrix, which contains the regression coefficients that appeared in the matrix model above, there are three new matrices in this equation:

- **A** a matrix that specifies weights we want to use on the independent variable coefficients.
- **C** a matrix that operates similarly on the dependent variable coefficients.
- **D** a matrix of null hypothesis values.

The apostrophe or prime mark (') after the **C** matrix denotes a transpose. It means that we flip over the **C** matrix so that its number of rows matches the number of columns in **B**.

MGLH allows you to input actual values for the **A**, **C**, and **D** matrices. You may never need to use them, because in almost every application, the values in them are either 1 or 0 and **MGLH** has automatic settings for these instances Their advantage, however, is that verbal descriptions can be ambiguous so that a statistician cannot be sure what hypothesis is being tested. When unusual hypotheses are tested, sometimes this is the only way to be sure that the test is specified correctly.

© 1989, SYSTAT, Inc.

Analysis of variance

Ordinarily, our dependent and independent variables are continuous. Often, however, we wish to examine the influence of categorical variables (such as gender, species, country, experimental group) on continuous variables. The model equations for this latter case, called *analysis of variance*, are the same as the ones we have been looking at. Somehow, however, we have to figure out a numerical coding for categories so we can use the codes in an equation as the independent variable(s). The usual method, used by **MGLH**, is called *effects coding*. In this method, each category is coded as a 1, 0, or –1.

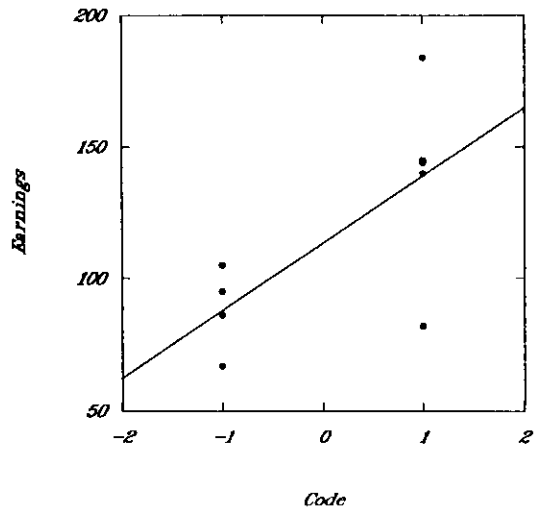
Two-group ANOVA

Here is an example. The following table shows the breakdown of our accountants sampled by sex. Since SEX is a categorical variable (numerical values assigned to MALE or FEMALE are arbitrary), we use by convention a CODE variable with the values 1 or –1. It doesn't matter which group is assigned –1 as long as the other is assigned 1. We have chosen –1 for females because the males earn more. Incidentally, by now you may be wondering whether we made up these data. We did, but the distributions on each variable closely follow data from an actual professional firm. Women partners tend to work comparable hours, have lower billings (perhaps because of limited opportunities to socialize with male corporate executive clients), and have lower earnings. The one variable we have modified is the sex distribution. In most firms, women partners are vastly outnumbered by males.

EARNINGS	SEX	CODE
86	female	-1
67	female	-1
95	female	-1
105	female	-1
86	female	-1
82	male	1
140	male	1
145	male	1
144	male	1
184	male	1

There is nothing wrong with plotting earnings against the code variable, as long as we realize that the slope of the line is arbitrary. By changing the values of the code variable, we can change the slope. Anyway, here is a plot with the least squares regression line superimposed.

© 1989, SYSTAT, Inc.



Now, the analysis of variance table in the **MGLH** output is still a global test of the significance of our model. The regression coefficients themselves are of less interest than the information in this table, however, because of the arbitrariness of the code variable scaling. The *F*-ratio in the Analysis of Variance table tells us that the coefficient for sex is significant at p=.019, which is less than the conventional .05 value. Thus, we can conclude on the basis of this sample and the validity of our usual regression assumptions, that women earn significantly less than men in this firm.

```
Dep var:EARNINGS      N:  10    Multiple R:  .719    Squared Multiple R:  .517

                        Analysis of Variance

Source      Sum-of-Squares    DF  Mean-Square     F-Ratio       P

SEX            6553.600        1    6553.600        8.563     0.019
Error          6122.800        8     765.350
```

The nice thing about realizing that ANOVA is specially coded regression is that the usual assumptions and diagnostics are appropriate in this context. You may plot residuals against estimated values, for example, to check for *homogeneity of variance* (constant error variance across categories).

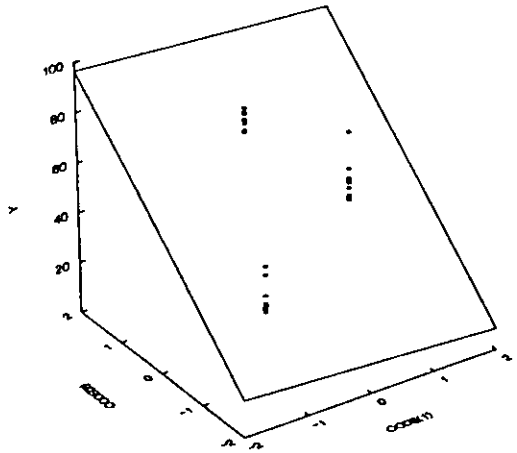
© 1989, SYSTAT, Inc.

Multi-group ANOVA

When there are more groups, the coding of categories becomes more complex. In general, there are one fewer coding variables than number of categories. For two categories, remember, we needed only one coding variable. For three categories, we need two:

Category	Code	
1	1	0
2	0	1
3	-1	-1

Here is a graph showing an ANOVA for three groups. The only difference from the 3-D plot of our multiple regression example above is that the data fall only in three distinct columns—one for each category.



You need not know how to produce coding variables in order to do ANOVA in SYSTAT. The program does all this for you automatically. When you use **MGLH**, all you need is a single variable that contains different values for each group. **MGLH** will translate these values into different codes. It is important to remember, however, that regression and analysis of variance are not fundamentally different models. They are both instances of the general linear model.

© 1989, SYSTAT, Inc.

Multivariate ANOVA

When you want to do ANOVA with more than one dependent variable, you can use the same type of model in **MGLH**:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

Now the **Y** matrix contains your dependent variables, the **X** matrix contains the codes for the ANOVA variable, and so on.

Conclusion

There are many other models **MGLH** can analyze. Factorial analysis of variance, for example, involves more than one categorical variable. Principal components and discriminant analysis are multivariate linear techniques. Because this is not a statistics book, however, there is not room here to discuss them in this section. If you need further information, consult the "Multivariate models: Examples" section at the end of this chapter. Each example contains additional introductory and theoretical material that can help you. You can also find further references in those sections.

© 1989, SYSTAT, Inc.

Computation

Missing data

MGLH eliminates any case with a missing value on any variable included in the model. If you want to do regression on a pairwise deletion covariance or correlation matrix, use **Corr** and the matrix input option of **MGLH**. This is a risky procedure, however, if you plan to test hypotheses.

Algorithms

The algorithms used in **MGLH** are all programmed in double precision (Wilkinson, 1975a, 1980). Centered sums of squares and cross products are accumulated using provisional algorithms (see the **Stats** chapter for references). Linear systems, including those involved in hypothesis testing, are solved via the modified Cholesky decomposition **LDL'**, where **L** is lower triangular and **D** is diagonal (no square roots). This procedure is rapid and accurate in double precision. Eigensystems are solved with Householder tridiagonalization and implicit QL iterations. For further information, see Wilkinson and Reinsch (1971) or Chambers (1977). The [Stepwise] option is implemented with forward and reverse sweeping (Dempster, 1969). Although this algorithm is used for ordinary regression calculations by most mainframe statistics packages, it can lead to round-off error with typical data, so it is not used in SYSTAT for calculation of coefficients and summary statistics. The user must estimate the subset model with the usual procedure. This means that regression coefficients and associated statistics printed by SYSTAT are generally more accurate than those printed by any widely used statistics package running on any computer.

© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

Computation

Missing data

MGLH eliminates any case with a missing value on any variable included in the model. If you want to do regression on a pairwise deletion covariance or correlation matrix, use Corr and the matrix input option of MGLH. This is a risky procedure, however, if you plan to test hypotheses.

Algorithms

The algorithms used in MGLH are all programmed in double precision (Wilkinson, 1975a, 1980). Centered sums of squares and cross products are accumulated using provisional algorithms (see the Stats chapter for references). Linear systems, including those involved in hypothesis testing, are solved via the modified Cholesky decomposition LDL', where L is lower triangular and D is diagonal (no square roots). This procedure is rapid and accurate in double precision. Eigensystems are solved with Householder tridiagonalization and implicit QL iterations. For further information, see Wilkinson and Reinsch (1971) or Chambers (1977). The [Stepwise] option is implemented with forward and reverse sweeping (Dempster, 1969). Although this algorithm is used for ordinary regression calculations by most mainframe statistics packages, it can lead to round-off error with typical data, so it is not used in SYSTAT for calculation of coefficients and summary statistics. The user must estimate the subset model with the usual procedure. This means that regression coefficients and associated statistics printed by SYSTAT are generally more accurate than those printed by any widely used statistics package running on any computer.

© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

Regression Contents

Introduction	143
Usage	145
Regression...	146
Saving residuals	146
Regression without the constant	146
General linear model...	147
Previewing your model	148
Saving residuals and other results	148
Cross validation	149
Durbin-Watson statistic	150
Regression without the constant	150
Repeated measures	150
Stepwise regression	150
Mixture models	150
Using an SSCP, covariance, or correlation matrix as input	150
Test...	152
Options	154
Selecting coefficients	155
Weighting coefficients	155
Show complete results	157
Examples	158
6.1 Simple regression	158
Saving residuals	160
6.2 Multiple regression	160
Extended output	162
Saving residuals	164
6.3 Testing hypotheses	165
Testing one regression coefficient	165
Testing two regression coefficients	166
Testing the same two coefficients with an A matrix	167
Testing a nonzero null hypothesis with a D matrix	168
6.4 Regression with ecological or grouped data	168
6.5 Weighted least squares	168
6.6 Regression without the constant	173
6.7 Mixture models	173
6.8 Polynomial regression	175
6.9 Stepwise regression	176
Options to control stepping	177
6.10 Partial correlations	178

© 1989, SYSTAT, Inc.

Regression Introduction

MGLH can estimate and test any univariate or multivariate general linear model. For an introduction to this broad class of models, see the "Introduction" to this chapter. Here in this "Regression" section, we focus on simple linear and multiple linear regression. For ANOVA, see the "Analysis of Variance" section, next, and for other multivariate models like discriminant analysis and principal components, see the final section of the chapter, "Multivariate models."

The model for linear regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

The matrix form of this model is:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

where y is a dependent variable vector, **X** is a vector or matrix of independent variable(s), β is a vector of regression coefficients, and ε is a vector of random errors. See Searle (1971), Neter and Wasserman (1985), or Cohen and Cohen (1983) for details.

Variations of this model include weighted least squares (e.g. using weights to stabilize the error variance across independent variables), mixture models (constraining the independent variables to sum to a constant), polynomial regression (including terms like x^2 and x^3 in the model), and stepwise regression (choosing a subset of predictors from among many candidates).

© 1989, SYSTAT, Inc.

Once the parameters of a model have been estimated, you can test that a coefficient (or a set of coefficients) is zero or test that a coefficient equals an hypothesized value. More generally, you can test any general linear hypothesis of the following form:

$$A\beta = D$$

where **A** is a matrix of linear weights on coefficients across the independent variables (the rows of β), β is the matrix of regression coefficients, and **D** is a null hypothesis matrix (usually a null matrix). These matrices (**A** and **D**) may be specified in several alternative ways and if they are not specified, they have default values.

For each model you fit, **MGLH** reports R^2, adjusted R^2, the standard error of the estimate, and an ANOVA table for assessing the fit of the model. For each variable in the model, the output includes the estimate of the regression coefficient, the standard error of the coefficient, and standardized coefficient, tolerance, and a *t* statistic for measuring the usefulness of the variable in the model.

You may also store results of the analysis (predicted values, residuals, and diagnostics that identify unusual cases) for further use in examining assumptions.

Most frequently your input will be the usual cases-by-variables data file. **MGLH** also can estimate regression models (but not compute residuals) directly from the covariance, correlation, or sums of squares and cross products matrix.

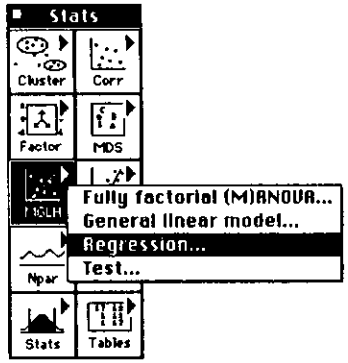
© 1989, SYSTAT, Inc.

Regression Usage

The **MGLH** submenu has two items for doing regression. The first item, **MGLH/Regression...**, is for simple and multiple regression models. The second, **MGLH/General linear model...**, is for more general regression models with dummy variables and interactions.

Regression... is easier to use than **General linear model...** when you are doing simple regression because the cursor automatically moves to the [Indep'dent] selection list after you've selected a dependent variable. Also, there are fewer options, so the dialog box is clean and simple. **Regression...** allows you to specify one dependent variable. All independent variables must be continuous. You can save residuals.

You must use **General linear model...** for including interaction terms in your model, for mixture models, and for stepwise regression. Only **General linear model...** lets you preview the equation form of the model you have set up.



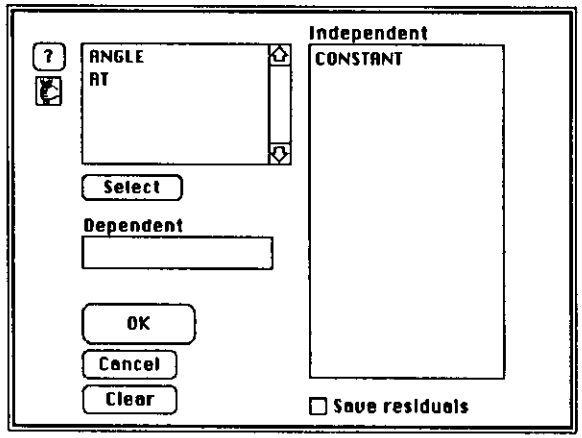
Press and hold the mouse button on **MGLH**. A submenu pops ups (like above). Select the regression item you want and release the mouse button. We show how to use each menu item below.

© 1989, SYSTAT, Inc.

Regression...

MGLH/Regression... lets you set up a simple regression model quickly and easily. After you select the dependent variable, the cursor automatically jumps to the [Indep'dent] selection box where you can select the independent variables.

- Select **MGLH/Regression...** from the **Stats** menu
- Select the dependent variable: RT
- Select the independent variable(s): ANGLE
- Click OK



Saving residuals

You can save residuals from any analysis by clicking [Save residuals]. A **Save as...** dialog box asks you to specify a filename, and after you click [Save], another dialog box lets you choose what gets saved and in what precision. For more information on this box or what gets saved, see "Saving residuals" in the **MGLH/General linear model...** section below.

Regression without the constant

You can delete the constant from your regression by selecting the CONSTANT term in the [Indep'dent] list and pressing the Delete key. You almost never want to remove the constant and you should be familiar with no-constant regression terminology before considering it. See also information about "Mixture models" below.

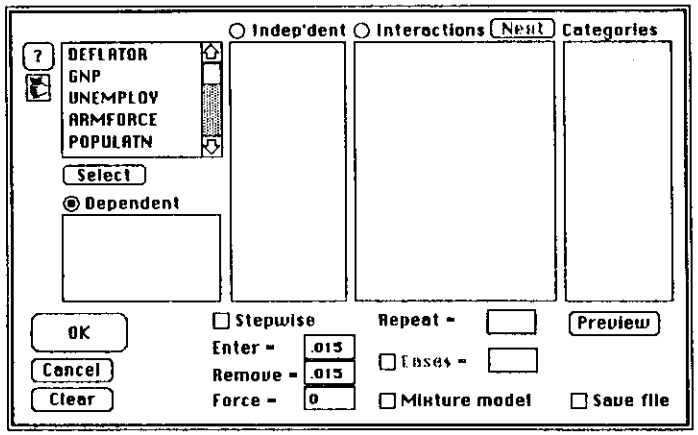
© 1989, SYSTAT, Inc.

General linear model...

MGLH/General linear model... is the most general item. You can specify any model with **General linear model...**, including those that you specify with **Fully factorial (M)ANOVA...** and **Regression...**, although ANOVA and simple regression models are more efficiently done with those items.

General linear model... has many options. You can ignore most of them. Only four steps—marked with solid bullets below—are required.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable(s)
- Click [Indep'dent] and select the independent variables
- Click [Interactions] and select the interaction terms, one at a time: for each interaction, select all the variables involved—SYSTAT links them with asterisks—and then click [Next] before selecting the variables involved in the next interaction term
- Click [Categories] and select those independent variables that are categorical; for each, SYSTAT asks you to specify the number of levels: type the number and click OK. SYSTAT will then generate a set of dummy variables (using effects coding) for each of the categorical variables you have specified. There will be one fewer dummy variables than the number of levels you specify for a given categorical variable.
- Click OK



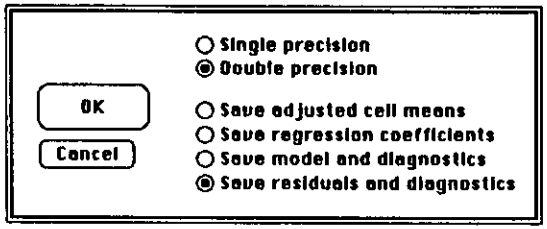
© 1989, SYSTAT, Inc.

Previewing your model

Click [Preview] to see your model displayed as an equation, with the dependent variables on the left and the independent variables (and any CONSTANT or interaction terms) on the right. Click OK to return to the **General linear model...** dialog box, where you can make any changes necessary. If you use the optional command interface, your MODEL statement is MODEL followed by the equation that you would see with [Preview].

Saving residuals and other results

You can save residuals from any analysis by clicking [Save file]. A **Save as...** dialog box asks you to specify a filename, and after you click [Save], another dialog box lets you choose what gets saved and in what precision:



Adjusted cell means. See Example 6.23, "Analysis of covariance" for more information about the contents of the file.

Regression coefficients. To save the coefficients instead of residuals, click [Save regression coefficients]. For multivariate models, coefficients for each dependent variable are saved into the file.

Model and diagnostics. If you click [Save model and diagnostics], the variables in your model are added to the residual and diagnostic variables in the file.

Residuals and diagnostics. The default option is [Save residuals and diagnostics]. The saved file contains three variables with the following labels:

ESTIMATE	estimated (predicted) values
RESIDUAL	residuals
LEVERAGE	leverage for each observation

© 1989, SYSTAT, Inc.

If your model has only one dependent variable, three additional variables are saved:

STUDENT	Studentized residuals
COOK	Cook's D values
SEPRED	Standard error of predicted values

Leverage is discussed in Belsley, Kuh, and Welsch (1980) and Velleman and Welsch (1981). Leverage helps you identify outliers in the independent variable space. As a rough rule of thumb for Gaussian data, leverage should have an average value of p/N, where p is the number of estimated parameters (including the constant) and N is the number of cases. An unusually large leverage might have a value greater than $2p/N$.

The Studentized residuals are the true "external" kind discussed in Velleman and Welsch (1981). Use these statistics to identify outliers in the dependent variable space. Under normal regression assumptions, they have a t distribution with $N–p$ degrees of freedom, where N is the total sample size and p is the number of predictors (including the constant).

Cook (1977) discusses the COOK statistic. COOK represents the influence of an observation on the estimates of the regression coefficients. Its formula combines leverage and Studentized residuals. Under the normal regression assumptions, COOK can be compared to an F distribution with p and $N–p$ degrees of freedom.

If there is more than one dependent variable (see examples below), the SYSTAT file contains an estimate and residual with subscripts for each one, e.g. ESTIMATE(1), ESTIMATE(2), RESIDUAL(1), RESIDUAL(2), LEVERAGE. The leverage is not subscripted. You must keep track of which variable corresponds to which subscript.

When you save residuals, SYSTAT prints warnings for large Cook's D statistics. These warnings appear in the output. The warning limits for the other statistics come from checking against t and F probability distributions.

Cross validation

If you weight the data before computing a model, **MGLH** computes predicted values for cases with zero weight, even though they are not included in the cases used to compute the regression equation.

© 1989, SYSTAT, Inc.

Durbin-Watson statistic

If you save residuals, the first order autocorrelation and Durbin-Watson statistic are displayed at the end of the output. See Example 6.5, "Weighted least squares." To compute additional autocorrelations or partial autocorrelations, use **Series/ACF plot...** or **Series/PACF plot...**

Regression without the constant

You can delete the constant from your regression by selecting the CONSTANT term in the [Indep'dent] list and pressing the Delete key. You almost never want to remove the constant and you should be familiar with no-constant regression terminology before considering it. See also information about "Mixture models," below.

Repeated measures

The [Repeat] option automatically creates contrasts on dependent variables in multivariate regression. You can use this capability for growth models and the analysis of multivariate profiles. This technique is most frequently used in the context of repeated measures analysis of variance. See the **Analysis of variance...** usage information below for details.

Stepwise regression

You can do stepwise selection of variables by clicking the [Stepwise] option. You can specify your own alpha-to-enter and alpha-to-remove values with [Enter] and [Remove] to control the stepping. You can also force the first *n* variables in your equation by specifying [Force=*n*]. The default values for these options are: Enter=.15, Remove=.15, Force=0. If your predictor variables are highly correlated, you might want to consider lowering the Enter and Remove values well below .05. See Example 6.9, "Stepwise regression," for more information about these options. If you delete the CONSTANT term from the model, [Stepwise] fits with intercept zero.

Mixture models

To test mixture models click the [Mixture model] option. With mixture models, the independent variables must sum to a constant value. You should not include a CONSTANT in the model. See Example 6.7, "Mixture models," for more information.

Using an SSCP, covariance, or correlation matrix as input: [Cases]

Normally, you analyze raw cases by variables data with **MGLH**. You can, however, use a symmetric matrix data file (for example, a covariance matrix saved in a file from **Corr**) as input. If you use a matrix as input, you *must* specify the sample size of the dataset that generated the matrix using the [Cases] option. The number you specify for [Cases] must be an integer greater than two.

© 1989, SYSTAT, Inc.

You can enter an SSCP, covariance, or correlation matrix by typing it into the Data Editor (see the *Getting Started* volume), by using DATA, or by saving it in a SYSTAT file and opening that file for **MGLH**. Be sure to include the dependent as well as independent variables in your matrix. **MGLH** picks out the dependent variable you name in your model.

SYSTAT uses the sample size to calculate degrees of freedom in hypothesis tests. **MGLH** determines the type of matrix (SSCP, covariance, etc.) and adjusts appropriately. With a correlation matrix, the raw and standardized coefficients are the same. Therefore, you cannot include a CONSTANT when using SSCP, covariance, or correlation matrices. Since these matrices are centered, the constant term has already been removed.

The following two analyses of the same data file produce identical results (except you don't get residuals with the second). In the first, we use the variables directly. In the second, we use **Corr** to save a covariance matrix and then analyze that matrix file with **MGLH**.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: Y
- Click [Indep'dent] and select the independent variables: X(1), X(2), X(3)
- Click OK

or

© 1989, SYSTAT, Inc.

- Select **Corr/Covariance...** from the **Stats** menu
- Select the variables: X(1), X(2), X(3), Y
- Click [Save file]
- Click OK

- Select **Open...** from the **File** menu
- Select the matrix file you saved
- Click Open

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: Y
- Click [Indep'dent] and select the independent variables: X(1), X(2), X(3)
- Delete the CONSTANT term
- Specify 40 for [Cases]
- Click OK

The triangular matrix input facility is useful for "meta-analysis" of published data and missing value computations. A few warnings, however. First, if you input correlation matrices from textbooks or articles, you may not get the same regression coefficients as those printed in the source. Because of round-off error, printed and raw data can lead to different results. Second, if you use pairwise deletion with **Corr**, the degrees of freedom for hypotheses will not be appropriate. You may not even be able to estimate the regression coefficients because of singularities.

In general, correlation matrices containing missing data produce regression coefficient estimates and hypothesis tests that are optimistic. You can correct for this by specifying a sample size smaller than the number of actual observations (preferably set it equal to the smallest number of cases used for any pair of variables), but this is a crude guess you could refine only by doing Monte Carlo simulations. There is no simple solution. Beware, especially, of multivariate regressions (or MANOVA, etc.) with missing data on the dependent variables. You can usually compute coefficients, but hypothesis testing is particularly suspect.

Test...

Test... lets you test parameters of the last model estimated. Linear hypotheses have the following form:

$$A\beta = D$$

© 1989, SYSTAT, Inc.

where **A** is a matrix of linear weights on coefficients across the independent variables (the rows of β), β is the matrix of regression coefficients, and **D** is a null hypothesis matrix (usually a null matrix). The **A** and **D** matrices may be specified in several alternative ways and if they are not specified, they have default values.

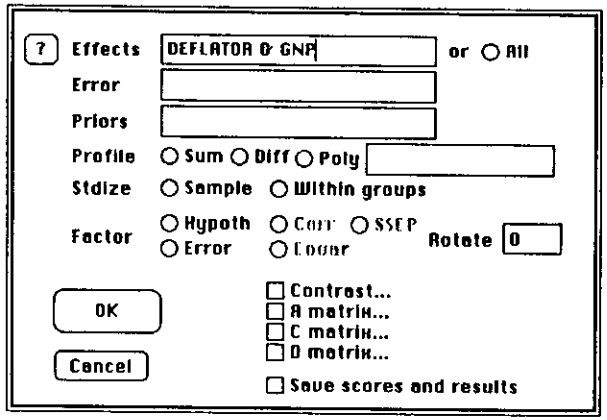
The ways to specify the **A** and **D** matrices are through the [Effects], [A matrix] and [D matrix] options in the **Test...** dialog box. In addition, the [C matrix] and [Profile] options are available for contrasting dependent variables in multivariate models and the [Contrast] option is for creating post-hoc contrasts in ANOVA. These are explained in more detail in the "Analysis of Variance" section.

For example, the **Test...** dialog box allows you to do things like the following:

1) If you want to test whether a single regression coefficient is zero, you can use the [Effects] box to specify the appropriate variable name. Of course, the *p* value for this test will be the same as that for the corresponding *t* statistic in the regression table.
2) If you want to test whether the coefficients for two variables simultaneously are zero, you can use the [Effects] box with ampersands (&) connecting the appropriate variable names. This test and the ones directly following do not appear in the regression output.
3) If you want to test whether two coefficients are equal, you can use the [A matrix] option to specify the contrast 1 vs. –1 on the appropriate variables.
4) If you want to test whether a coefficient is three times larger than another, you can use the [A matrix] option to specify the contrast 1 vs. –3 on the appropriate variables.
5) If you want to test whether one coefficient is equal to 30 and another equal to 50, you can identify the appropriate variables in the [Effects] box (separated by ampersands) and specify the hypothesized values 30 and 50 in the [D matrix] .
6) If you want to test whether the difference between two coefficients is 30 , you can use [A matrix] to specify the appropriate 1 vs. –1 contrast and [D matrix] to specify 30.

© 1989, SYSTAT, Inc.

- Select **MGLH/Test...** from the **Stats** menu
- Specify [Effects]; use asterisks to indicate interactions; use ampersands to test several effects simultaneously (see below)
- Click OK



Options

The following seven options are more relevant to Analysis of Variance and other multivariate models:

[Error]
Specify an error term other than residual mean square. See the "Analysis of Variance" section.

[Priors]
Specify prior probabilities for discriminant analysis. See the "Multivariate models" section.

[Stdize]
Standardize canonical coefficients using total sample or within groups covariance matrix. See the "Multivariate models" section.

[Profile]
Contrast dependent variables with sum, difference, or polynomial weights for repeated measures and growth models. See the "Analysis of Variance" section.

© 1989, SYSTAT, Inc.

[Factor]
Compute principal components on the hypothesis (between groups) or error (within groups) correlation, covariance, or SSCP matrix. See the "Multivariate models" section.

[Rotate]
Rotate canonical factors. See the "Multivariate models" section.

[Save scores and results]
Save MANOVA or canonical variable (discriminant) scores and predicted group membership into a file. See the "Analysis of variance" and "Multivariate models" sections.

Selecting coefficients: [Effects]

The easiest way to specify one or more regression coefficients for testing in a hypothesis is to use the [Effects] option. Simply type the name of the variable whose coefficient you want to test. If you have more than one variable in mind, then you must separate each variable name with an ampersand (&), e.g.:

GNP&DEFLATOR

If you wish to test an interaction term, then you must separate each variable name with an asterisk (*), e.g.

GNP*DEFLATOR*TIME

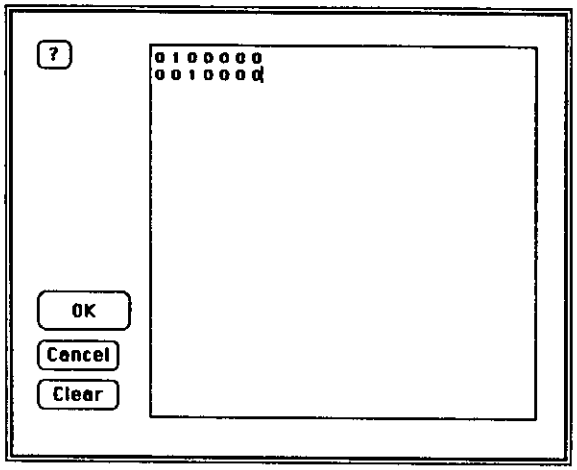
You may need to use both asterisks and ampersands. Suppose you wanted to test the effects of A, B, and the interaction of A and B. You would type "A & B & A*B" in the [Effects] box.

The [All] option is a shortcut if you wish to test all the effects in your model. [All] automatically generates separate single degree of freedom hypotheses. The resulting *p* values are the same as the ones for the t tests of significance of the coefficients in the regression output. This option is handy when using the [C matrix] or [Profile] options in multivariate models, however.

Weighting coefficients: [A matrix] and [D matrix]

If you need to weight coefficients to test hypotheses, the matrix options allow you to enter specific numerical weights. Remember, the [Effects] box amounts to using weights whose values are restricted to 1 or 0. The matrix options allow you to input the A and D matrices directly. Both the [A matrix] and [D matrix] options bring up similar dialog boxes:

© 1989, SYSTAT, Inc.



The **A** matrix has as many columns as there are regression coefficients (including the constant) in your model. The number of rows in **A** determine how many degrees of freedom your hypothesis involves. The **D** matrix, if you use it, must have the same number of rows as **A**. For multiple regression, **D** has only one column. For multivariate models (multiple dependent variables), the **D** matrix has one column for each dependent variable. See Example 6.3 for **A** and **D** matrix applications in regression.

- Select **MGLH/Test...** from the **Stats** menu
- Click [A matrix] and type the matrix, separating columns with spaces; press Return between rows
- Click OK

Use the same procedure to enter a **D** matrix.

The matrix dialog boxes retain the matrices you specify until you enter new ones, click [Clear], or close the data file. However, the default (an identity matrix) is used for subsequent tests unless you reopen the same box and click [OK] before testing another hypothesis.

© 1989, SYSTAT, Inc.

Show complete results

As with other items, **MGLH** produces extended output if you choose [Show extended results] from **Data/Formats...** The default is [Show standard results]. You get extended output from many of the routines until you change the setting in **Data/Formats...**

For regression, [Show extended results] adds the following output: eigenvalues of the unit scaled matrix **X'X**, condition indices, variance proportions, and a correlation matrix of regression coefficients. These are discussed in Example 6.2, "Multiple regression," under the subheading "Extended output."

For hypothesis testing, [Show extended results] adds **A**, **C**, and **D** matrices, the matrix of contrasts, and the inverse of the cross products of contrasts. For multivariate regression, you also get hypothesis and error sum of product matrices, tests of residual roots, canonical correlations, coefficients, and loadings.

The inverse of the contrasts basis matrix is $\mathbf{A(X'X)^{-1}A'}$. The latter contrast can be used for constructing Scheffé or simultaneous confidence intervals on regression coefficients or contrasts of coefficients (Morrison, 1976). SYSTAT does not compute the inverse of **X'X** to solve for regression coefficients because this introduces inaccuracies. You can display it, however, by letting the **A** matrix default to the identity for **Test...**

© 1989, SYSTAT, Inc.

Regression Examples

You should read this section sequentially, because the models are presented in order of increasing complexity. Information on interpreting a given part of **MGLH** output is presented with the first example that generates that output. If you skip ahead and don't know what part of the output means, look for explanations in earlier examples.

6.1 Simple regression

The ROTATE dataset from Metzler and Shepard (1974) records reaction time versus angle of rotation in a perception study. Shepard had shown in a number of previous experiments that mental images were used in higher cognitive processes. The experiment was one of several that measured the time in milliseconds it took subjects to make "same" judgments when comparing a picture of a three dimensional object to a picture of possible rotations of the object.

ANGLE	RT
2	1.8
22	2.4
40	3.0
62	2.7
85	4.0
100	4.1
130	5.0
140	4.8
163	4.9
180	5.7

We want to find the slope of the best fitting least squares line relating reaction time to angle of rotation. Here is how to do this in **MGLH**:

- Select **MGLH/Regression...** from the **Stats** menu
- Select the dependent variable: RT
- Select the independent variable: ANGLE
- Click OK

Notice how the model sentence specifies the simple linear equation with a constant and slope term.

© 1989, SYSTAT, Inc.


```
Dep Var:      RT      N:   10    Multiple R:  .974    Squared Multiple R: . 949
Adjusted Squared Multiple R:  .943     Standard Error of Estimate:        0.310

  Variable    Coefficient    Std Error     Std Coef Tolerance     T     P(2 Tail)

CONSTANT          1.916         0.185         0.000    .          10.358    0.000
   ANGLE          0.021         0.002         0.974  1.000        12.261    0.000

                              Analysis of Variance

   Source    Sum-of-Squares    DF  Mean-Square     F-Ratio       P

Regression        14.417        1       14.417     150.326      0.000
  Residual         0.767        8        0.096
```

In this example, the Multiple R (.974) is the simple Pearson correlation because there is only one predictor in the equation. The Squared Multiple R (.949) denotes the proportion of variance in the dependent variable accounted for by the predictor ANGLE in this linear model. The Adjusted Squared Multiple R (.943) reduces this proportion to a level expected when using this model in a new sample from the same population. The formula for this adjusted multiple correlation statistic is

$$1 - \frac{(1-R^2)(n-1)}{(n-p)}$$

where n is the number of cases and p is the number of predictors, including the CONSTANT. Finally, the Standard Error of Estimate (.310) is the square root of the residual mean square (.096) at the bottom of the table.

The regression coefficients for predicting RT are found in the next part of the table (CONSTANT=1.92, ANGLE=.021). The standard errors of these coefficients are in the next column and the standardized coefficients (called "betas" by some social scientists) are in the next. Tolerance is not relevant in this context; it is one minus the squared multiple correlation between each predictor and the remaining predictors in the equation. Since there is only one predictor (ANGLE) in the equation, it is 1. In multiple regressions, as in the next example, values of Tolerance near zero indicate that some of the predictors are highly intercorrelated, an undesirable situation in regression.

© 1989, SYSTAT, Inc.

There are two *t* statistics reported in the table. The first (10.36) tests the significance of the difference of the constant from zero. This is not meaningful here since we assume a minimum delay in responding is physically inevitable. The second (12.26) tests the second coefficient, which is equivalent to testing the significance of the correlation between RT and ANGLE.

The Analysis of Variance table is really overkill here. The F-ratio (150.326) is the square of the *t*-statistic (12.26) for the coefficient. This table is more useful for assessing multiple regression models.

Saving residuals

You should usually click [Save residuals] in the **Regression...** dialog box and click [Save model and diagnostics] for simple regression so that you can plot residuals against the independent variable or estimated values. You can do a variety of graphics by opening the new residuals file. In this example, it will contain RT, ANGLE, ESTIMATE, RESIDUAL, LEVERAGE, STUDENT, COOK, and SEPRED. See the "Usage" section above for definitions of these variables. If you plot RESIDUAL or STUDENT against ESTIMATE or ANGLE, you will see that the distribution of residuals is well behaved. This reflects Shepard's care as an experimenter and expertise as a statistician.

6.2 Multiple regression

This example is a multiple regression using a data set (LONGLEY) from Longley (1967). This data set was originally used to test the robustness of least squares packages to multicollinearity and other sources of ill-conditioning. **MGLH** can print more "correct" digits than the solution provided by Longley himself if you select 9 [Decimals to be shown] in **Data/Formats...**, although the default limits the printed precision to three digits after the decimal.

© 1989, SYSTAT, Inc.

DEFLATOR	GNP	UNEMPLOY	ARMFORCE	POPULATN	TIME	TOTAL
83.0	234289	2356	1590	107608	1947	60323
88.5	259426	2325	1456	108632	1948	61122
88.2	258054	3682	1616	109773	1949	60171
89.5	284599	3351	1650	110929	1950	61187
96.2	328975	2099	3099	112075	1951	63221
98.1	346999	1932	3594	113270	1952	63639
99.0	365385	1870	3547	115094	1953	64989
100.0	363112	3578	3350	116219	1954	63761
101.2	397469	2904	3048	117388	1955	66019
104.6	419180	2822	2857	118734	1956	67857
108.4	442769	2936	2798	120445	1957	68169
110.8	444546	4681	2637	121950	1958	66513
112.6	482704	3813	2552	123366	1959	68655
114.2	502601	3931	2514	125368	1960	69564
115.7	518173	4806	2572	127852	1961	69331
116.9	554894	4007	2827	130081	1962	70551

- Select **MGLH/Regression...** from the **Stats** menu
- Select the dependent variable: TOTAL
- Select the independent variables: DEFLATOR, GNP, UNEMPLOY, ARMFORCE, POPULATN, TIME
- Click OK

```
Dep Var:   TOTAL      N:   16    Multiple R:  .998    Squared Multiple R:  .995
Adjusted Squared Multiple R:  .992     Standard Error of Estimate:      304.854

  Variable     Coefficient     Std Error      Std Coef Tolerance     T     P(2 Tail)

CONSTANT      -3482258.635   890420.384          0.000     .       -3.911    0.004
DEFLATOR          15.062       84.915          0.046    .007      0.177    0.863
     GNP          -0.036        0.033         -1.014    .001     -1.070    0.313
UNEMPLOY          -2.020        0.488         -0.538    .030     -4.136    0.003
ARMFORCE          -1.033        0.214         -0.205    .279     -4.822    0.001
POPULATN          -0.051        0.226         -0.101    .003     -0.226    0.826
    TIME        1829.151      455.478          2.480    .001      4.016    0.003

                        Analysis of Variance

   Source    Sum-of-Squares    DF  Mean-Square     F-Ratio       P

Regression    .184172E+09      6  .306954E+08      330.285      0.000
  Residual    836424.056       9    92936.006
```

Notice the extremely small tolerances in the output. These signal that the predictor variables are highly intercorrelated—a worrisome situation. This *multicollinearity* can inflate the standard errors of the coefficients, thereby attenuating the associated *t* statistics, and can even threaten computational accuracy.

After the output is displayed, you can use **Regression...** again or test any hypotheses involving linear combinations of regression coefficients.

© 1989, SYSTAT, Inc.

Extended output

Extended regression output is available when you choose [Show complete results] from **Data/Formats...** Reestimate the model:

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Click OK

- Select **Redo last analysis** from the **Goodies** menu

or specify the model again manually:

- Select **MGLH/Regression...** from the **Stats** menu
- Select the dependent variable: TOTAL
- Select the independent variables: DEFLATOR, GNP, UNEMPLOY, ARMFORCE, POPULATN, TIME
- Click OK

This produces the following additional output between the regression coefficients and the analysis of variance table printed above:

```
Eigenvalues of Unit Scaled X'X

          1          2          3          4          5

      6.861      0.082      0.046      0.011      0.000

          6          7

      0.000      0.000

Condition Indices

          1          2          3          4          5

      1.000      9.142     12.256     25.337    230.424

          6          7

   1048.080  43275.042
```


© 1989, SYSTAT, Inc.


```
Variance Proportions

                    1          2          3          4          5

CONSTANT        0.000      0.000      0.000      0.000      0.000
DEFLATOR        0.000      0.000      0.000      0.000      0.457
     GNP        0.000      0.000      0.000      0.001      0.016
UNEMPLOY        0.000      0.014      0.001      0.065      0.006
ARMFORCE        0.000      0.092      0.064      0.427      0.115
POPULATN        0.000      0.000      0.000      0.000      0.010
    TIME        0.000      0.000      0.000      0.000      0.000

                    6          7

CONSTANT        0.000      1.000
DEFLATOR        0.505      0.038
     GNP        0.328      0.655
UNEMPLOY        0.225      0.689
ARMFORCE        0.000      0.302
POPULATN        0.831      0.160
    TIME        0.000      1.000

Correlation Matrix of Regression Coefficients

                CONSTANT   DEFLATOR        GNP   UNEMPLOY   ARMFORCE

CONSTANT           1.000
DEFLATOR          -0.205      1.000
     GNP           0.816     -0.649      1.000
UNEMPLOY           0.836     -0.555      0.946      1.000
ARMFORCE           0.550     -0.349      0.469      0.619      1.000
POPULATN          -0.411      0.659     -0.833     -0.758     -0.189
    TIME          -1.000      0.186     -0.802     -0.824     -0.549

                POPULATN       TIME

POPULATN           1.000
    TIME           0.388      1.000
```

SYSTAT computes these statistics by scaling the columns of the **X** matrix so that the diagonal elements of **X'X** are ones and then factoring the **X'X** matrix . In the Longley example most of the eigenvalues of **X'X** are nearly zero, showing that the predictor variables comprise a relatively redundant set. The Condition Indices are the square roots of the ratios of the largest eigenvalue to each successive eigenvalue. A condition index greater than 15 indicates a possible problem and one greater than 30 suggests a serious problem with collinearity (Belsley, Kuh, and Welsh, p. 198). The condition indices in the Longley example show a tremendous collinearity problem.

© 1989, SYSTAT, Inc.

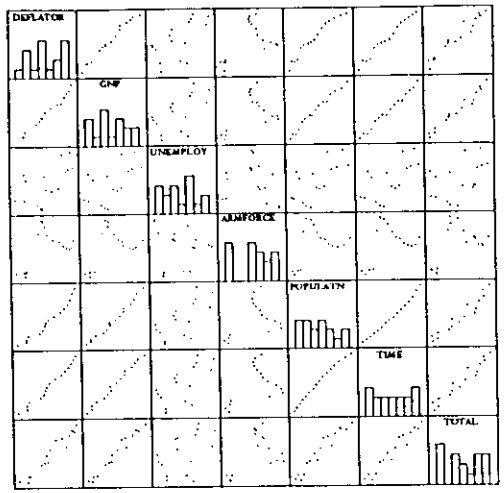
The Variance Proportions are the proportion of the variance of the estimate accounted for by each principal component associated with each of the above eigenvalues. You should begin to worry about collinearity when a component associated with a high condition index contributes substantially to the variance of two or more variables. This is certainly the case with the last component of the Longley data. TIME, GNP, and UNEMPLOY load highly on this component. See Belsley, Kuh, and Welsch (1980) for more information on these diagnostics.

Finally, SYSTAT gives the Correlation Matrix of Regression Coefficients. In the Longley data, these estimates are highly correlated, further indicating that there too many predictors in the equation to provide stable estimates.

Saving residuals

You should usually click [Save residuals] in the **MGLH/Regression...** dialog box and click [Save model and diagnostics] for regression so that you can plot residuals against the independent variables and estimated values. You can do a variety of graphics by opening the new residuals file. In this example, it will contain all the Longley variables, as well as ESTIMATE, RESIDUAL, LEVERAGE, STUDENT, COOK, and SEPRED. See the "Usage" section above for definitions of these variables. In addition, you should examine a scatterplot matrix of all the Longley variables. Notice in the figure below the severely nonlinear distributions of ARMFORCE with the other variables, as well as the near perfect correlations among several of the predictors. There is also a sharp discontinuity between post-war and 1950's behavior on ARMFORCE.

© 1989, SYSTAT, Inc.



6.3 Testing hypotheses

You can test single regression coefficients, several coefficients, linear combinations of coefficients, and nonzero null hypotheses with the [Effects], [A matrix], and [D matrix] options in the **MGLH/Test...** dialog box.

Testing one regression coefficient

Most regression programs print tests of significance for each coefficient in an equation. **MGLH** has a powerful additional feature: post-hoc tests of regression coefficients. To demonstrate these tests, we examine whether the DEFLATOR coefficient differs significantly from zero. This is equivalent to saying H_0: $\beta_1 = 0$.

- Select **MGLH/Test...** from the **Stats** menu
- Specify DEFLATOR for [Effects]
- Click OK

```
Test for effect called:
                DEFLATOR

Test of Hypothesis

    Source          SS       DF          MS          F          P

Hypothesis     2923.976        1     2923.976      0.031      0.863
     Error   836424.056        9    92936.006
```

© 1989, SYSTAT, Inc.

Notice the Error Sum of Squares (836424.056) is the same as the Residual Sum of Squares at the bottom of the ANOVA table. The probability level (.863) is the same also.

You can test all the coefficients in the equation this way, individually, or you click [All] for the [Effects] to generate separate hypothesis tests for each predictor (independent variable) you used for the regression.

- Select **MGLH/Test...** from the **Stats** menu
- Click [All] next to [Effects]
- Click OK

Testing two regression coefficients

You may wonder why you need to bother with this **Test...** stuff when the regression output gives you the same numbers. Try this one. It tests the following: H_0: $\beta_1 = \beta_2 = 0$.

- Select **MGLH/Test...** from the **Stats** menu
- Specify DEFLATOR & GNP for [Effects]; be sure to use the ampersand between effects
- Click OK

```
Test for effect called:
                DEFLATOR
           and
                GNP

A Matrix

                    1          2          3          4          5

          1     0.000      1.000      0.000      0.000      0.000
          2     0.000      0.000      1.000      0.000      0.000

                    6          7

          1     0.000      0.000
          2     0.000      0.000

Test of Hypothesis

     Source         SS       DF         MS          F          P

 Hypothesis  149295.592        2  74647.796      0.803      0.478
      Error  836424.056        9  92936.006
```

This time, the Error Sum of Squares is the same as before, but the Hypothesis Sum of Squares is different. We just tested the hypothesis that the DEFLATOR and GNP coefficients simultaneously are zero.

© 1989, SYSTAT, Inc.

The A Matrix printed above the test specifies the hypothesis we tested. It has two degrees of freedom (see the *F*-statistic) because the **A** matrix has two rows—one for each coefficient. If you know some matrix algebra, you can see that the matrix product **AB** using this **A** matrix and **B** as a column matrix of regression coefficients picks up only two coefficients: DEFLATOR and GNP. Notice that our hypothesis had the following matrix equation: **AB** = **0**, where **0** is a null matrix.

If you don't know matrix algebra, don't worry; the method of specifying variables for [Effects] in the **Test...** dialog box is equivalent. You can ignore the A Matrix in the output.

Testing the same two coefficients with an A matrix

If you are a matrix algebra freak, however, you can specify your own matrix with [A matrix] in the **Test...** dialog box. Be sure to separate cells with spaces and press Return between rows. This is the same null hypothesis as before: H_0: $\beta_1 = \beta_2 = 0$.

- Select **MGLH/Test...** from the **Stats** menu
- Click [A matrix] and enter the matrix:
  0 1 0 0 0 0 0
  0 0 1 0 0 0 0
- Click OK

You will get the same output as above.

Why bother with **A matrix...** when the [Effects] option does the same job? Because in the **A** matrix you may use any numbers, not just zeroes and ones. Here is a bizarre matrix:

1.0 3.0 0.5 64.3 3.0 2.0 0.0

You may not want to test this kind of hypothesis on the Longley data, but there are important applications in the analysis of variance where you might. We will see some of these later in the chapter.

© 1989, SYSTAT, Inc.

Testing a nonzero null hypothesis with a D matrix

You can test nonzero null hypotheses with a **D** matrix, often in combination with [Effects] or [A matrix]. Here, we test whether the DEFLATOR coefficient significantly differs from 30. It tests the following: H_0: $\beta_1 = 30$.

- Select **MGLH/Test...** from the **Stats** menu
- Click [A matrix] and enter the matrix:
  0 1 0 0 0 0 0
- Click [D matrix] and enter the matrix:
  30
- Click OK

6.4 Regression with ecological or grouped data

If you have aggregated data, weight the regression by a count variable. This variable should represent the counts of observations (n_i) contributing to each "case." If n is not an integer, SYSTAT truncates it to an integer before using it as a weight. The regression results are identical to those produced if you had typed in each case n_i times.

Here is an example using the PLANTS dataset:

CO2	SPECIES	COUNT
12	1	10
14	2	9
13	3	15
11	5	22
11	6	20

- Select **Weight...** from the **Data** menu
- Select a weight variable: COUNT
- Click OK

- Select **MGLH/Regression...** from the **Stats** menu
- Select the dependent variable: CO2
- Select the independent variable: SPECIES
- Click OK

The total N for this regression is 76.

6.5 Weighted least squares

The following data are from Neter, Wasserman, and Kutner (1985). We add three variables to this data set (WEIGHT, WX, and WY) to demonstrate weighted least squares. Here we input the data in the command window. You can do this stuff with the Data Editor if you prefer.

© 1989, SYSTAT, Inc.


```
DATA
SAVE NIELSEN
INPUT X Y
LET WEIGHT=SQR(1/X^2)
LET WX=X*WEIGHT
LET WY=Y*WEIGHT
RUN
  2.13    15.5
  1.21    11.1
 11.00    62.6
  6.00    35.4
  5.60    24.9
  6.91    28.1
  2.97    15.0
  3.35    23.2
 10.39    42.0
  1.10    10.0
  4.36    20.0
  8.00    47.5
~
```

Before doing weighted least squares, let's try ordinary least squares:

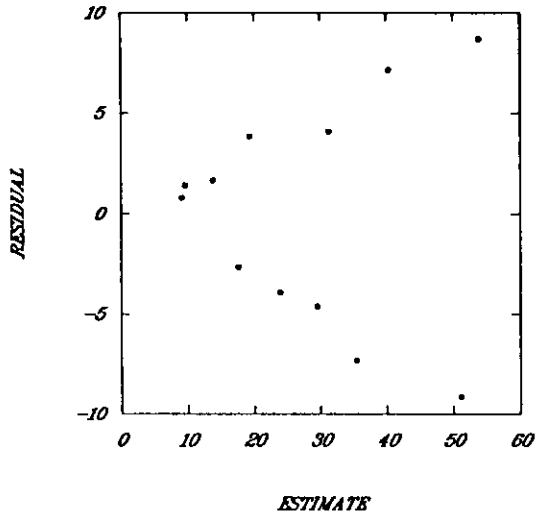
- Select **MGLH/Regression...** from the **Stats** menu
- Select the dependent variable: Y
- Select the independent variable: X
- Click [Save residuals]
- Click OK

```
Dep Var:       Y      N:   12    Multiple R:  .938    Squared Multiple R:  .880
Adjusted Squared Multiple R:  .867     Standard Error of Estimate:          5.870

  Variable    Coefficient    Std Error     Std Coef Tolerance    T    P(2 Tail)

CONSTANT            4.229        3.252        0.000 .          1.301    0.223
       X            4.515        0.528        0.938 1.0000000  8.544    0.000

                        Analysis of Variance

   Source    Sum-of-Squares    DF  Mean-Square     F-Ratio       P

 Regression       2515.113     1     2515.113      73.000       0.000
   Residual        344.536    10       34.454

Durbin-Watson D Statistic    1.840
First Order Autocorrelation   .002

Residuals have been saved
```

© 1989, SYSTAT, Inc.

We saved the residuals so we could plot them.

- Select **Plot/Plot** from the **Graph** menu
- Select the Y variable: RESIDUAL
- Select the X variable: ESTIMATE
- Click OK



Notice the fan shaped, or heteroscedastic, distribution of the residuals. This violates the assumption, needed for hypothesis tests, of homogeneity of variance in the residuals across different values of X. Weighted least squares often helps with this kind of problem. A weighting variable controls the influence of different cases on the fitted function. Cases at values of X with larger error variance are forced to have less influence than other cases.

In this example, the error variance is assumed to be proportional to X^2, so our weights are the reciprocal of this value. Notice in our DATA step above, however, that we used the square root of these weights for our weighting variable. Now we produce the weighted estimates:

© 1989, SYSTAT, Inc.

- Select **MGLH/Regression...** from the **Stats** menu
- Select the dependent variable: WY
- Select the independent variable: WEIGHT, WX
- Delete the CONSTANT term from the independent variable list
- Click [Save residuals]
- Click OK

We have multiplied all the variables in the equation (including the constant) by the square root of our weight.

```
Model contains no constant

Dep Var:      WY      N:   12   Multiple R:  .992   Squared Multiple R:  .983
Adjusted Squared Multiple R:  .981   Standard Error of Estimate:          0.890

  Variable    Coefficient    Std Error    Std Coef Tolerance      T    P(2 Tail)

  WEIGHT          5.657        0.965        0.378  .4047928     5.861     0.000
      WX          4.191        0.404        0.670  .4047928    10.381     0.000

                          Analysis of Variance

  Source    Sum-of-Squares    DF  Mean-Square     F-Ratio       P

Regression       461.447      2      230.724      291.499     0.000
Residual           7.915     10        0.792

Durbin-Watson D Statistic     1.855
First Order Autocorrelation  -.007

Residuals have been saved
```

The output tells us that the model has no constant; this is because we used our weighting variable as a "constant." The multiple correlation, therefore, is inflated because it includes variance about absolute zero rather than the mean of Y. For the same reason, the Regression Sum of Squares (461.447) has 2 degrees-of-freedom instead of 1. To adjust the regression sum of squares for our "constant" term, do the following:

- Select **MGLH/Test...** from the **Stats** menu
- Specify WX for [Effects]
- Click OK

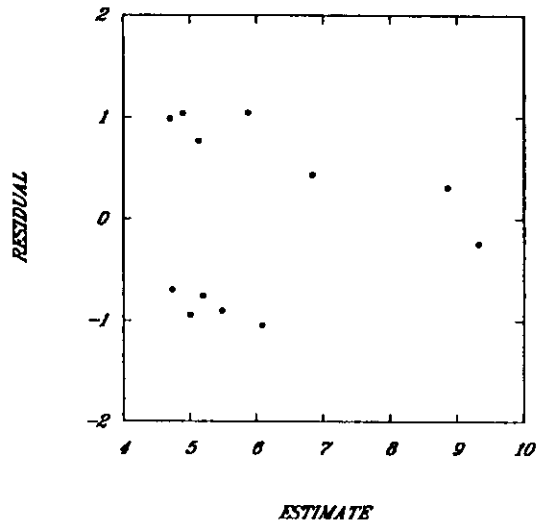
The sum of squares for this hypothesis is the "adjusted" regression sum of squares.

Notice that the estimates of the coefficients are slightly different, since the cases have been weighted differently. Furthermore, the standard errors of the coefficients are considerably smaller, improving our inference.

© 1989, SYSTAT, Inc.

Again, we plot the residuals.

- Select **Plot** from the **Graph** menu
- Select the Y variable: RESIDUAL
- Select the X variable: ESTIMATE
- Click OK



Remember to use the square root of whatever weights you choose to get the correct weighted estimates. You can consult Neter, Wasserman, and Kutner (1985) or Draper and Smith (1981) for more information.

© 1989, SYSTAT, Inc.

6.6 Regression without the constant

To regress without the constant (intercept) term, or regression through the origin, remove the CONSTANT from the list of independent variables. Just select CONSTANT and press the Delete key. **MGLH** adjusts accordingly.

Some users are puzzled when they see a model without a constant having a higher multiple correlation than a model that includes a constant. How can a regression with fewer parameters predict "better" than another? It doesn't. The total sum of squares must be redefined for a regression model with zero intercept. It is no longer centered about the mean of the dependent variable. Other definitions of sums of squares can lead to strange results such as negative multiple correlations. If your constant really is near zero, then including or excluding the constant makes little difference in the output. Kvålseth (1985) discusses the issues involved in summary statistics for zero-intercept regression models. The definition of R^2 used in SYSTAT is Kvålseth's formula number 7. This was chosen because it retains its PRE (percentage reduction of error) interpretation and is guaranteed to be in the (0,1) interval.

How, then, do you test the significance of a constant in a regression model? Include CONSTANT in the model statement as usual and look at its test of significance. In Example 6.2 with the Longley data, the *t* statistic for the constant is -3.91.

If you have a zero-intercept model where it is appropriate to compute a coefficient of determination and other summary statistics about the centered data, use the [Mixture model] option of **General linear model...** This option analyzes mixture models and is discussed in the next section (Example 6.7). It provides Kvålseth's formula 1 for R^2 and uses centered total sum of squares for other summary statistics.

6.7 Mixture models

Mixture models decompose the effects of mixtures of variables on a dependent variable. They differ from ordinary regression models because the independent variables sum to a constant value. The regression model, therefore, does not include a constant and the regression and error sums of squares have one less degree of freedom. Marquardt and Snee (1974) and Diamond (1981) discuss these models and their estimation.

© 1989, SYSTAT, Inc.

Here is an example using the PUNCH dataset from Cornell (1985). The study involved effects of various mixtures of watermelon, pineapple, and orange juice on taste ratings by judges of a fruit punch.

W	P	O	VALUE
1.0	0.0	0.0	4.3
1.0	0.0	0.0	4.7
1.0	0.0	0.0	4.8
0.0	1.0	0.0	6.2
0.0	1.0	0.0	6.5
0.0	1.0	0.0	6.3
0.5	0.5	0.0	6.3
0.5	0.5	0.0	6.1
0.5	0.5	0.0	5.8
0.0	0.0	1.0	7.0
0.0	0.0	1.0	6.9
0.0	0.0	1.0	7.4
0.5	0.0	0.5	6.1
0.5	0.0	0.5	6.5
0.5	0.0	0.5	5.9
0.0	0.5	0.5	6.2
0.0	0.5	0.5	6.1
0.0	0.5	0.5	6.2

Use the [Mixture model] option to analyze these data:

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the dependent variable: VALUE
- Click [Indep'dent] and select the independent variables: W, P, O
- Click [Interactions] and select the interactions by selecting pairs of variables, clicking [Next] between pairs: W*P [Next], W*O [Next], P*O
- Click [Mixture model]
- Click OK

```
Model contains no constant.

Assuming mixture model.

Dep Var:   VALUE      N:   18    Multiple R:  .969    Squared Multiple R:  .939
Adjusted Squared Multiple R:  .913     Standard Error of Estimate:        0.232

  Variable    Coefficient    Std Error     Std Coef Tolerance    T    P(2 Tail)

       W          4.600        0.134        3.001 0.6666667   34.322    0.000
       P          6.333        0.134        4.131 0.6666667   47.255    0.000
       O          7.100        0.134        4.631 0.6666667   52.975    0.000
      W*
       P          2.400        0.657        0.320 0.6666667    3.655    0.003
      W*
       O          1.267        0.657        0.169 0.6666667    1.929    0.078
      P*
       O         -2.200        0.657       -0.293 0.6666667   -3.351    0.006
```


© 1989, SYSTAT, Inc.


```
Analysis of Variance

  Source    Sum-of-Squares    DF  Mean-Square     F-Ratio       P

Regression         9.929     5        1.986       36.852     0.000
 Residual          0.647    12        0.054
```

Not using the [Mixture model] option produces a much larger R^2 (.999) and an F-value of 2083.371, both of which are inappropriate for these data. Notice the Regression Sum-of-Squares has 5 degrees of freedom instead of 6 as in the usual zero-intercept regression model. We have lost one degree-of-freedom because the predictors sum to 1.

6.8 Polynomial regression

MGLH can fit curvilinear functions with polynomials. This is useful for response surface methodology in designed experiments. For observational studies, however, it is often a sleazy method, because it is easy to overfit the data and polynomials rarely have simple theoretical meaning. It is usually better to use **Nonlin** to fit a theoretical function you have devised yourself. Let's try a simple example from the **Nonlin** chapter that uses the USPOP (U.S. population) dataset. We will fit the data with a curve made up of linear and quadratic components.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: POP
- Click [Indep'dent] and select the independent variable: YEAR
- Click [Interaction] and select YEAR twice for the YEAR*YEAR interaction
- Click OK

```
Dep Var:     POP      N:   19    Multiple R:  .999    Squared Multiple R:  .998
Adjusted Squared Multiple R:  .998     Standard Error of Estimate:     2781.025

  Variable    Coefficient    Std Error     Std Coef Tolerance    T    P(2 Tail)

CONSTANT      .204504E+08   843475.326        0.000 .          24.245    0.000
    YEAR      -22780.606      897.849      -20.280 .0001683  -25.372    0.000
   YEAR*
    YEAR          6.346        0.239       21.242 .0001683   26.576    0.000

                        Analysis of Variance

  Source    Sum-of-Squares    DF  Mean-Square     F-Ratio       P

Regression    .717990E+11     2  .358995E+11    4641.719     0.000
 Residual     .123746E+09    16  7734097.817
```

© 1989, SYSTAT, Inc.

The Residual Sum of Squares (.123746E+09) is one-tenth of the figure produced by **Nonlin** (.126348E+10) with the same number of parameters. The equation estimated in **Nonlin** is based on a theoretical model of exponential growth. The polynomial equation above has no theoretical basis. I do not think, however, that theoretical models should necessarily be rejected in favor of atheoretical models that fit better. Goodness-of-fit should be the servant of theory, not its master.

Incidentally, we can improve our fit even more by adding a YEAR*YEAR*YEAR interaction. We could even add a fourth degree polynomial. *Ad nauseam.*

6.9 Stepwise regression

Use the [Stepwise] option of **General linear model...** to compute a stepwise regression. Here is an example using the Longley data.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: TOTAL
- Click [Indep'dent] and select the independent variables: DEFLATOR, GNP, UNEMPLOY, ARMFORCE, POPULATN, TIME
- Click the [Stepwise] option
- Click OK

```
Stepwise Regression with Alpha-to-enter= .150 and Alpha-to-remove= .150


Step= 1     Enter      GNP     R= .984     Rsquare= .967
Step= 2     Enter UNEMPLOY     R= .990     Rsquare= .981
Step= 3     Enter ARMFORCE     R= .993     Rsquare= .985
Step= 4     Enter     TIME     R= .998     Rsquare= .995

The subset model includes the following predictors:

        CONSTANT
             GNP
        UNEMPLOY
        ARMFORCE
            TIME

Use these predictors in a new model sentence to estimate the coefficients
```

SYSTAT selected four predictors in four steps. None of the predictors were removed from the equation, resulting in a final equation with a constant and four predictors. If you leave CONSTANT out of the equation, [Stepwise] fits with zero intercept.

© 1989, SYSTAT, Inc.

Notice that only information on the multiple correlation is printed. If you want full information on any step, estimate a new model with that predictor using **General linear model...** or **Regression....** SYSTAT uses this method because algorithms for stepwise regression can introduce round-off error when many steps are necessary, particularly when predictors are highly intercorrelated. Even though the stepping is done in double precision, it is safer to reestimate the model.

Options to control stepping

Several options control the stepping. Specify the alpha-to-enter with [Enter], and alpha-to-remove with [Remove].

Do not make alpha-to-remove less than alpha-to-enter, or you may cycle variables in and out of the equation (stepping stops automatically if this happens). The standard values of alpha (.15) were chosen on the basis of Monte Carlo studies of stepwise regression (Bendel & Afifi, 1977). If you want to force some variables into the equation before stepping (they will never be removed), use the [Force] option—specify the number of variables you want forced. For instance, setting [Force] to 3 forces the first three variables in your most recent model into the equation.

If you are thinking of testing hypotheses with **Test...** after fitting a subset model, don't bother. Stepwise regression programs are the most notorious source of pseudo "*p*-values" in the field of automated data analysis. Stepwise regression programs that print pseudo *F*-tests and *p*-values are inviting abuse; statisticians seem to be the only ones who know these are not "real" *p*-values. The stepwise option is provided in **MGLH** to select a subset model for prediction purposes. It should never be used without cross validation.

If you still want some sort of confidence estimate on your subset model, you might look at tables in Wilkinson (1979), Rencher and Pun (1980), and Wilkinson and Dallal (1982). These tables provide null hypothesis R^2 values for selected subsets given the number of candidate predictors and final subset size. If you don't know this literature already, you will be surprised at how large multiple correlations from stepwise regressions on random data can be. For a general summary of these and other problems, see Hocking (1983). For more specific discussions of variable selection problems see the previous references plus Flack and Chang (1987), Freedman (1983), and Lovell (1983).

© 1989, SYSTAT, Inc.

It is not difficult to write a stepwise regression program that looks like the ones in other popular statistics packages. The reason we will not put such a program in SYSTAT is that stepwise regression is probably the most abused computerized statistical technique ever devised. If you think you need stepwise regression to solve a particular problem, it is almost certain that you do not. Professional statisticians rarely use automated stepwise regression because it does not necessarily find (a) the "best" fitting model, (b) the "real" model, or (c) alternative "plausible" models. Furthermore, the order in which variables enter or leave a stepwise program is usually of no theoretical significance. You are always better off thinking about why a model could generate your data and then testing that model.

6.10 Partial correlations

Partial correlations are easy to compute with **MGLH**. The partial correlation of two variables (A and B) given a third (C) is the correlation between the residuals of each (A and B) after each has been regressed on the third (C). You can therefore use **MGLH** to compute an entire matrix of partial correlations.

Here is an example of how you would compute a matrix of partial correlations of Y(1–5) given X:

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Click OK

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: Y(1), Y(2), Y(3), Y(4), Y(5)
- Click [Indep'dent] and select the independent variable: X
- Click OK

Look for the "Residual Correlation Matrix" in the output; it is the matrix of partial correlations among the Y's given X. If you want to partial several X's, just select them (also) as independent variables.

© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

Analysis of variance Contents

Introduction 181
Usage 182
Fully factorial (M)ANOVA... 183
Saving residuals 184
Repeated measures 184
General linear model... 184
Options 185
Previewing your model 186
Repeated measures 186
Saving residuals and other results 186
Durbin-Watson statistic 188
ANOVA without the constant 188
Test... 188
Options 190
Selecting effects for hypothesis tests 190
Comparisons and contrasts 191
Selecting error terms 193
[Profile] 194
Show complete results 194
Examples 195
6.11 Analysis of variance 195
6.12 Two-way analysis of variance 202
6.13 Single degree of freedom designs 212
6.14 Unbalanced designs 216
6.15 Randomized block designs 219
6.16 Incomplete block designs 220
6.17 Fractional factorial designs 221
6.18 Mixed models 223
6.19 Nested designs 225
6.20 Split plot designs 227
6.21 Latin square designs 230
6.22 Crossover and changeover designs 232
6.23 Analysis of covariance 234
6.24 Hotelling's T^2 237
6.25 Multivariate analysis of variance 238
6.26 Multivariate analysis of covariance 244
6.27 Multivariate profile analysis 245
6.28 Unweighting means 250
6.29 One-way repeated measures 254
6.30 Repeated measures with grouping factors 256
6.31 Repeated measures with two trial factors 264
6.32 Repeated measures with three or more trials factors 270
6.33 Repeated measures analysis of covariance 274
6.34 Covariance alternatives to repeated measures 277
6.35 Issues in repeated measures analysis 277

© 1989, SYSTAT, Inc.

Analysis of variance Introduction

MGLH can handle a wide variety of balanced and unbalanced analysis of variance designs. It is a true least squares program, so unbalanced designs are handled by least squares estimates rather than adjusted harmonic means. The model specification encompasses fully crossed designs, randomized blocks, incomplete blocks, fractional factorials, and Latin squares. For models with fixed and random effects, you can define error terms for specific hypotheses. For a simple one-way ANOVA with post-hoc tests of pairwise mean differences, see **Stats**.

MGLH also includes repeated measures, split plot, and crossover designs. It includes both univariate and multivariate approaches to repeated measures designs.

Once you have estimated your ANOVA model, it is easy to test post-hoc pairwise differences in means or to test any contrast across cell means, including simple effects.

You may store results of the analysis (predicted values and residuals) for further study and graphical display. In ANCOVA, you may save adjusted cell means. See Example 6.23.

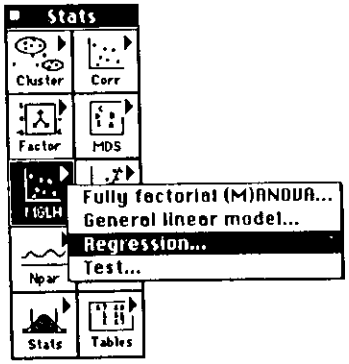
© 1989, SYSTAT, Inc.

Analysis of variance Usage

MGLH has two items for doing ANOVA. The first, **Fully factorial (M)ANOVA...**, does factorial analysis of variance and covariance. The second, **MGLH/General Linear Model...**, gives you more control over your model, so that you can fit incomplete factorials, Latin squares, and nested designs.

Fully factorial (M)ANOVA... requires that all independent variables (factors) are categorical. All interactions are included in the model and tested automatically. You can specify covariates, do repeated measures, and save residuals. **Fully factorial (M)ANOVA...** does not display effect estimates (regression coefficients) unless you choose [Show complete results] from **Data/Formats...** before estimating the model.

Fully factorial (M)ANOVA... is easier to use than **General linear model...** (when the factors are all categorical and interactions should all be tested) because you only need to select the factors once, and you immediately specify the number of levels for each factor. With **General linear model...**, you must specify all the interactions, whereas **Fully factorial (M)ANOVA...** generates and tests them automatically.



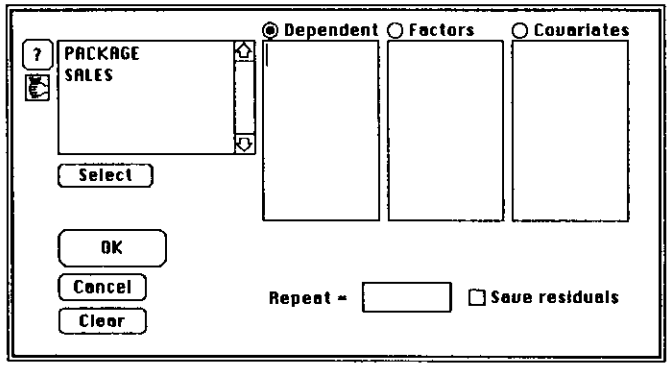
Press and hold the mouse button on **MGLH**. A submenu pops ups. Select the ANOVA item you want and release the mouse button. We show how to use each menu item in the following sections.

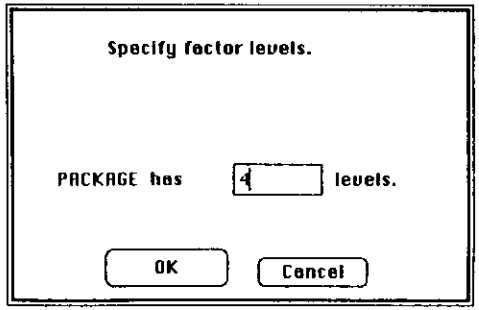
© 1989, SYSTAT, Inc.

Fully factorial (M)ANOVA...

MGLH/Fully factorial (M)ANOVA... does analysis of variance and multivariate analysis of variance for fully factorial models—that is, models where all the independent variables are categorical variables with values 1, 2, ..., *k*, where *k* is the number of groups.

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variable(s)
- Click [Factors] and select the factors
  - A dialog box asks you to specify factor levels; type the number of levels and click OK
- Click OK





© 1989, SYSTAT, Inc.

The [Dependent] list contains your dependent variables (within cells measures). The [Factors] list contains your categorical grouping variables that define the cells. The [Covariates] list contains optional covariates, which are continuous concomitant measures.

Saving residuals

You can save residuals from any analysis by clicking [Save residuals]. A **Save as...** dialog box asks you to specify a filename, and after you click [Save], another dialog box lets you choose what gets saved and in what precision. For more information on this dialog box or what gets saved, see "Saving residuals" in the **MGLH/General linear model...** section below.

Repeated measures

To get repeated measures hypothesis tests, specify the number of levels for each repeated measures factor. Your measures should be included in the [Dependent] window. Each trial measurement should appear in the list as a dependent variable.

If you have one repeated measures factor with 3 levels (trials), for example, you should have three variables in the [Dependent] list and a "3" in the [Repeat] box. If you have two repeated factors with 3 and 4 levels respectively, you should have 12 variables in the [Dependent] window and "3,4" in the [Repeat] box. In this latter instance, the variables in the selection list should be ordered as 3 blocks of 4 variables each. In the first block, the four levels would be the ordered trials on the second factor, and so on. You may not have more than two repeated measures factors.

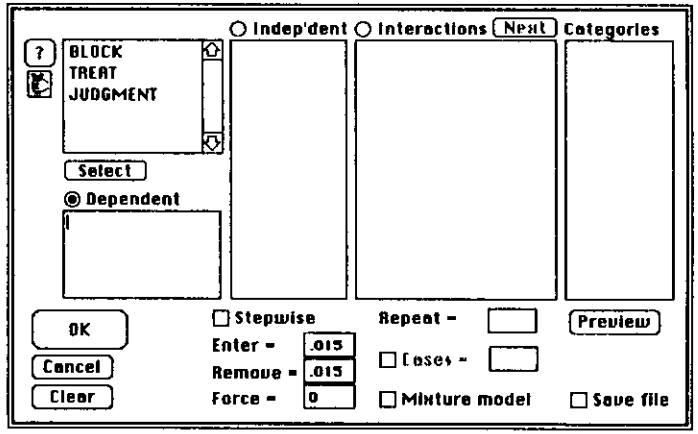
General linear model...

MGLH/General linear model... is the most general item. You can specify any model with **General linear model...**, including those that you specify with **Fully factorial (M)ANOVA...** and **Regression...**, although ANOVA and simple regression models are more efficiently done with those items.

General linear model... has many options. You can ignore most of them. Only four steps—marked with solid bullets below—are required.

© 1989, SYSTAT, Inc.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable(s)
- Click [Indep'dent] and select the independent variables
- Click [Interactions] and select the interaction terms, one at a time: for each interaction, select all the variables involved—SYSTAT links them with asterisks—and then click [Next] before selecting the variables involved in the next interaction term
- Click [Categories] and select those independent variables that are categorical; for each, SYSTAT asks you to specify the number of levels: type the number and click OK. SYSTAT will then generate a set of dummy variables (using effects coding) for each of the categorical variables you have specified. There will be one fewer dummy variables than the number of levels you specify for a given categorical variable.
- Click OK



Options

The following three options are more relevant to regression and other multivariate models:

[Stepwise]
Compute a stepwise regression, with [Enter] and [Remove] values, and [Force] criterion. See the "Regression" section.

© 1989, SYSTAT, Inc.

[Cases]
Defines sample size for alternative matrix input for regression and other multivariate models. See the "Regression" and "Multivariate models" sections.

[Mixture model]
Omit constant and treat predictors as mixtures that sum to a constant. See the "Regression" section.

Previewing your model

Click [Preview] to see your model displayed as an equation, with the dependent variables on the left and the independent variables (and any CONSTANT or interaction terms) on the right. Click OK to return to the **General linear model...** dialog box, where you can make any changes necessary. If you use the optional command interface, your MODEL statement is MODEL followed by the equation that you would see with [Preview].

Repeated measures

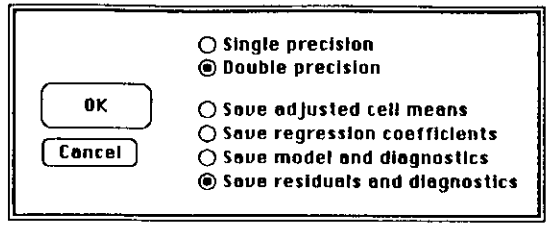
For repeated measures designs, specify the number of levels for each repeated measures factor. Your measures should be specified in the [Dependent] window. Each trial measurement should appear in the list as a dependent variable.

If you have one repeated measures factor with 3 levels (trials), for example, you should have three variables in the [Dependent] window and a "3" in the [Repeat] box. If you have two repeated factors with 3 and 4 levels respectively, you should have 12 variables in the [Dependent] window and "3,4" in the [Repeat] box. In this latter instance, the variables in the selection list should be ordered as 3 blocks of 4 variables each. In the first block, the four levels would be the ordered trials on the second factor, and so on. You may not have more than two repeated measures factors. See the examples (6.29–33) to see how to set up the model. You will need to rearrange your data file a bit if you have more than one factor.

Saving residuals and other results

You can save residuals from any analysis by clicking [Save file]. A **Save As** dialog box asks you to specify a filename, and after you click [Save], another dialog box lets you choose what gets saved and in what precision:

© 1989, SYSTAT, Inc.



Adjusted cell means. These are least squares estimates of cell means adjusted for covariates. See Example 6.23, "Analysis of covariance" for more information about the contents of the file.

Regression coefficients. These are estimates of ANOVA effects. To save them instead of residuals, click [Save regression coefficients]. For multivariate models, coefficients for each dependent variable are saved into the file.

Model and diagnostics. If you click [Save model and diagnostics], the (effects coded) dummy variables in your model are added to the residual and diagnostic variables in the file. Dummy variables, which comprise the design matrix, are labeled with subscripts.

Residuals and diagnostics. The default option is [Save residuals and diagnostics]. The saved file contains three variables with the following labels:

ESTIMATE	estimated cell means
RESIDUAL	residuals
LEVERAGE	leverage for each observation

If your model has only one dependent variable, three additional variables are saved:

STUDENT	Studentized residuals
COOK	Cook's D values
SEPRED	Standard error of predicted values

Leverage is discussed in Belsley, Kuh, and Welsch (1980) and Velleman and Welsch (1981). For balanced designs (equal numbers of cases per cell), leverage is not of interest. With unbalanced designs, unusually sparse cells will have large leverage values.

© 1989, SYSTAT, Inc.

The Studentized residuals are the true "external" kind discussed in Velleman and Welsch (1981). Like leverage, this statistic depends on the type of dummy variable coding, and so should be interpreted cautiously. See Section 1, "Regression," for more discussion of their usefulness.

Cook (1977) discusses the COOK statistic. Like leverage and Studentized residuals, this statistic depends on the type of dummy variable coding.

If there is more than one dependent variable (see examples below), the SYSTAT file contains an estimate and residual with subscripts for each one, e.g. ESTIMATE(1), ESTIMATE(2), RESIDUAL(1), RESIDUAL(2), LEVERAGE. The leverage is not subscripted. You must keep track of which variable corresponds to which subscript.

Durbin-Watson statistic

If you save residuals, the first order autocorrelation and Durbin-Watson statistic are displayed at the end of the output. (See Example 6.5, "Weighted least squares.") To compute additional autocorrelations or partial autocorrelations, use **Series/ACF plot...** or **Series/PACF plot...** In ANOVA, the value of this statistic depends on the experimental design and the order in which the data appear in the file. You should interpret it cautiously, but a large value can signal danger.

ANOVA without the constant

You can delete the constant from your ANOVA by selecting the CONSTANT term in the Independent window and pressing the Delete key. You almost never want to remove the constant. ANOVA models without a constant are extremely rare.

Test...

Test... lets you test parameters of the ANOVA model estimated. Linear hypotheses have the following form:

$$\mathbf{ABC'} = \mathbf{D}$$

where **A** is a matrix of linear weights contrasting the ANOVA effects (the rows of **B**), and **B** is the matrix of effects, with one column for univariate designs and several columns for MANOVA. **C** is a matrix of contrasts on the dependent variables for MANOVA designs. **C** has as many columns as there are dependent variables. For ANOVA, **C** is simply the number 1. Finally, **D** is a null hypothesis matrix (usually a null matrix). These matrices (**A**, **C**, and **D**) may be specified in several alternative ways, and if they are not specified they have default values.

© 1989, SYSTAT, Inc.

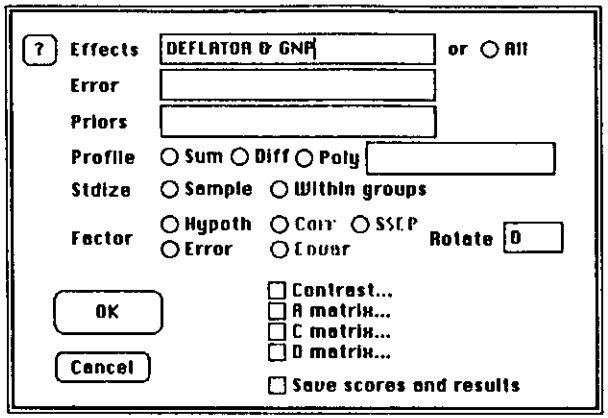
You can specify the **A**, **C**, and **D** matrices with the [Effects], [A matrix], [C matrix], [D matrix], [Contrast], and [Profile] options in the **Test...** dialog box.

For example, the **Test...** dialog box allows you to do things like:

1) If you want to test whether a factor is significant, you can use the [Effects] option to specify the name of the factor. This test, of course, duplicates the *F*-test you can find in the ANOVA table.
2) If you want to test whether the effects of two factors simultaneously are zero, you can use the [Effects] option with ampersands (&) connecting the appropriate variable names. This test cannot be found in the ANOVA table. It is a type of omnibus *F*-test for the significance of several factors taken together.
3) If you want to test whether two marginal means are equal, you can use [Effects] to identify the factor and [Contrast] to specify the appropriate contrast coefficients (1 and –1) on the levels of this factor.
4) If you want to test whether two cell means are equal, you can use the [A matrix] option to specify the appropriate contrast.
5) If you want to test whether the difference between two cell means is 30 , you can use [A matrix] to specify the appropriate 1 vs. –1 contrast as in (4) and the [D matrix] to specify 30.

- Select **MGLH/Test...** from the **Stats** menu
- Specify [Effects]; use asterisks to indicate interactions; use ampersands to test several effects simultaneously (see below)
- Click OK

© 1989, SYSTAT, Inc.



Options

The following four options are more relevant to Regression and other multivariate models:

[Priors]
Specify prior probabilities for discriminant analysis. See the "Multivariate models" section.

[Stdize]
Standardize canonical coefficients using total sample or within groups covariance matrix. See the "Multivariate models" section.

[Factor]
Compute principal components on the hypothesis (between groups) or error (within groups) correlation, covariance, or SSCP matrix. See the "Multivariate models" section.

[Rotate]
Rotate canonical factors. This option can be used for MANOVA to reveal a simple structure in the canonical variate space. See the "Multivariate models" section. See Wilkinson (1977) for an example.

Selecting effects for hypothesis tests: [Effects]

The easiest way to specify one or more effects for testing in a hypothesis is to use the [Effects] box. Type the name of the effect you want to test. If you have more than one effect in mind, then you must separate each effect name with an ampersand (&), e.g.

© 1989, SYSTAT, Inc.

GROUP&REGION

If you wish to test an interaction term, separate each variable name with an asterisk (*), e.g.

SEX*DRUG*GROUP

You may need to use both asterisks and ampersands. Suppose you wanted to test the effects A, B, and the interaction of A and B. You would type "A & B & A*B" in the [Effects] box.

The [All] button is a shortcut if you wish to test all the effects in your model. [All] automatically generates separate tests for each factor and interaction term in the model. The resulting *p* values are the same as the ones for the *F* tests of significance in the ANOVA table. This option is handy when using the [C matrix] or [Profile] options for multivariate models.

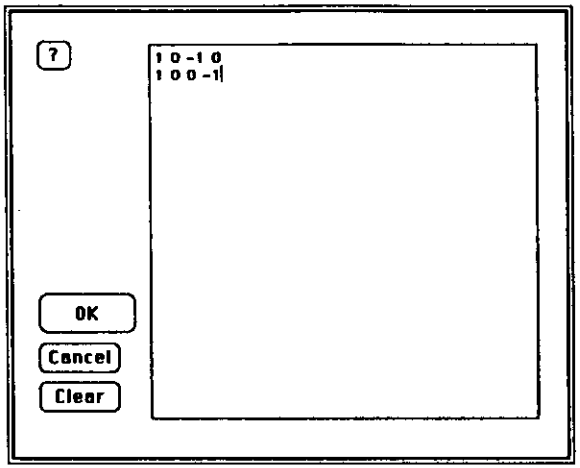
Comparisons and contrasts: [Contrast], [A matrix], [C matrix], [D matrix]

If you need to compare or contrast effects or cell means, there are several different methods ranging from contrasts within factors using the [Contrast] option, to general linear contrasts using the matrix options.

Contrasts within factors

The [Contrast] option can be used together with [Effects] to specify a contrast among the levels of a factor. You cannot click [Contrast] until you have specified a factor name within [Effects]. See Examples 6.11 and 6.12 ("Analysis of variance" and "Two-way analysis of variance") for introductory information on contrast coding. Look for the subheading "Post-hoc tests." Further information appears in subsequent examples.

© 1989, SYSTAT, Inc.



- Specify a factor name in the [Effects] box
- Click [Contrasts]
- Type the contrast matrix, separating columns with spaces; press Return between rows. Each row is a single degree of freedom contrast.
- Click OK

[Contrast] retains the matrix you specify until you enter a new one, click [Clear], or close the data file.

General linear contrasts

You can enter general linear contrasts the same way you enter [Contrast]. See the [Contrast] example directly above. Clicking [A matrix], [C matrix], or [D matrix] brings up a dialog box for you to enter a matrix.

© 1989, SYSTAT, Inc.

The **A** matrix has as many columns as there are degrees of freedom (including the grand mean) in your model. The number of rows in **A** determines how many degrees of freedom your hypothesis involves. Remember, the [Effects] option amounts to using weights whose values are restricted to 1 or 0. The [A matrix] option allows you to input any real valued coefficients. Example 6.11 shows how [A matrix] and [Contrast] correspond. See the end of Example 6.12 for how to use an A matrix to test simple effects.

The **C** matrix has as many columns as there are dependent variables, or measures. The number of rows in **C** helps determine the degrees of freedom in multivariate tests. Although the general linear contrast is of the form **ABC'=D**, note that you should enter **C**, not **C'**. When entering the matrix, it is easier to think in terms of columns for your dependent variables than rows. See Example 6.27 for a discussion of **C** matrices.

The **D** matrix, if you use it, must have the same number of rows as **A**. For ANOVA, **D** has only one column. For MANOVA, the **D** matrix has one column for each dependent variable. Usually, you test the null hypothesis where **D=0**, a null matrix. This is the default. You can specify your own **D** to test any null hypothesis, however.

Selecting error terms: [Error]

Normally, **MGLH** uses the residual error (within cells) for testing hypotheses. To use another source from your model, specify it for [Error]. Simply type the name of the error term you want to specify. This term must be one of the terms in your model, which you can see if you click [Preview] in the **MGLH/General Linear Model...** dialog box.

If you have more than one effect in mind, then you must separate each effect name with an ampersand (&), e.g.

A&B

If you wish to test an interaction term, then you must separate each variable name with an asterisk (*), e.g.

A*B*C

You can use both asterisks and ampersands to pool error terms. Suppose you wanted the effects A, B, and the interaction of A and B. You would type "A & B & A*B" for [Error].

© 1989, SYSTAT, Inc.

See Example 6.18, "Mixed models" and Example 6.20, "Split plot designs" for more information.

[Profile]

The [Profile] options are most often used for repeated measures designs. They create several different types of C matrix directly. Choose [Sum] to compute and test the sum of dependent variables, [Diff] to compute pairwise differences of successive measures, and [Poly] to generate orthogonal polynomials. For [Poly], you may enter your own metric to produce unequally spaced polynomials. If you check [Poly] and enter nothing in the box beside it, you will get orthogonal polynomials across equally spaced intervals. For more information on profile analysis, see Example 6.27, "Multivariate profile analysis."

Show complete results

As with other items, **MGLH** produces extended output if you choose [Show extended results] from **Data/Formats...** The default is [Show standard results]. You get extended output from many of the routines until you change the setting in **Data/Formats...**

For analysis of variance, [Show extended results] adds the following: estimates of ANOVA effects (regression coefficients) and coefficients of determination (squared multiple correlations). See Example 6.11, "Analysis of variance" and Example 6.25, "Multivariate Analysis of Variance."

© 1989, SYSTAT, Inc.

Analysis of variance Examples

6.11 Analysis of variance

Let's start with a simple example: one-way (one factor) analysis of variance. The following dataset, KENTON, is from Neter, Wasserman, and Kutner (1985). The data comprise unit sales of a cereal product under different types of package designs. The replications of PACKAGE represent different stores.

PACKAGE	SALES
1	12
1	18
2	14
2	12
2	13
3	19
3	17
3	21
4	24
4	30

Package designs have been randomly assigned to stores, although each design does not have an equal number of stores. Notice that we do not put sales in four different columns. Instead, we use the grouping variable PACKAGE to identify the package type. SALES, after all, is the same variable (the dependent variable) in all the cells.

- Select **MGLH/Fully factorial (M)ANOVA...** from the Stats menu
- Select the [Dependent] variable: SALES
- Click [Factors] and select the factor: PACKAGE
  - Specify the number of levels for PACKAGE: 4
- Click OK

© 1989, SYSTAT, Inc.

Any variable selected in the [Factors] list must have integer values ranging from 1 to *n* where *n* is the number of levels for that variable. The value for each case of PACKAGE, therefore, must be 1, 2, 3, or 4. **MGLH** uses these values to index the different levels of PACKAGE.

```
Dep Var:  SALES     N:  10   Multiple R:  .921   Squared Multiple R:  .849

                    Analysis of Variance

  Source   Sum-of-Squares   DF  Mean-Square   F-Ratio      P

 PACKAGE          258.000    3       86.000    11.217   0.007

   Error           46.000    6        7.667
```

This is the standard analysis of variance table. The *F*-Ratio (11.217) appears significant, so we could conclude that the package designs differ significantly in their effects on sales, provided the experimental assumptions are valid.

Effects and dummy variables

You could do the same analysis of variance above with **General linear model...** by selecting SALES as the dependent variable and PACKAGE as an independent variable. The only difference is that **Fully factorial (M)ANOVA...** does not display regression coefficients, because you generally want to see the sums of squares associated with ANOVA effects rather than the effects themselves. If you do want to see the estimates of the effects when you use **Fully factorial (M)ANOVA...**, choose [Show extended results] from **Data/Formats...** before you do the analysis.

The effects in a least squares analysis of variance are associated with a set of "dummy" variables that **MGLH** generates automatically. Ordinarily, you do not have to concern yourself with these dummy variables. If you want to see them, however, you can save them into a SYSTAT file by clicking [Save file] and choosing [Save model and diagnostics] in the dialog box that greets you after you name the file. This is the same option that saves residuals, estimates, model variables, and other regression statistics.

© 1989, SYSTAT, Inc.

To view the file, just open it in the Data Editor. Here are the data:

SALES	X(1)	X(2)	X(3)
12	1	0	0
18	1	0	0
14	0	1	0
12	0	1	0
13	0	1	0
19	0	0	1
17	0	0	1
21	0	0	1
24	-1	-1	-1
30	-1	-1	-1

The variables X(1–3) are the dummy variables generated by **MGLH** in response to your selecting a factor and specifying a number of levels. All cases in the first cell are associated with dummy values 1 0 0, those in the second cell with 0 1 0. The third, 0 0 1, and the fourth, –1 –1 –1. Other least squares programs use different methods to code dummy variables. The coding used by **MGLH** is most widely used and guarantees that the effects sum to zero.

The file we saved contains residuals and other diagnostic statistics. We don't ordinarily think of these as being relevant to ANOVA, but they are critically important. Years ago, testing of ANOVA assumptions was confined to looking at Bartlett's test for homogeneity of variance (see **Stats**). People thought ANOVA was robust to violations of assumptions. Daniel (1960) and others have shown us how wrong that attitude was. One case can influence an entire ANOVA. We will use some of these diagnostic statistics in our two-way ANOVA example.

Post-hoc tests

Once you have estimated your ANOVA model, it is easy to do post-hoc tests in **MGLH**. This example compares the first and third packages:

- Select **MGLH/Test...** from the **Stats** menu
- Specify PACKAGE for [Effects]
- Click [Contrast] and enter the matrix:
  1 0 –1 0
- Click OK

© 1989, SYSTAT, Inc.

We specified only one effect, so our model has one degree of freedom; therefore, the contrast matrix has only one row of numbers. These numbers are the same ones you see in ANOVA textbooks, though **MGLH** offers one convenience: you do not have to standardize them so that their sum of squares are 1. Just type integers.

Here is the output:

```
Test for effect called:
                PACKAGE

A Matrix

                1         2         3         4
            0.000     1.000     0.000    -1.000

Test of Hypothesis

    Source        SS       DF        MS         F         P

Hypothesis    19.200        1    19.200     2.504     0.165
     Error    46.000        6     7.667
```

The *F* statistic (2.504) is not significant, so we cannot conclude that the impact of the first and third package designs on sales is significantly different. Incidentally, the A matrix printed contains the contrast. The first column (0) corresponds to the constant in the model and the remaining three columns (1 0 –1) correspond to the dummy variables for PACKAGE. If this doesn't make any sense to you, don't worry. You don't have to understand it to compute contrasts.

Most ANOVA programs allow you to do pairwise comparisons like this. **MGLH** allows any linear contrast. You can, for example, compare multiple treatments with a control. Let's compare the average performance of the first three packages with the last:

- Select **MGLH/Test...** from the **Stats** menu
- Specify PACKAGE for [Effects]
- Click [Contrast] and enter the matrix:
  1 1 1 –3
- Click OK

© 1989, SYSTAT, Inc.


```
Test for effect called:
                PACKAGE

A Matrix

                1         2         3         4
            0.000     4.000     4.000     4.000

Test of Hypothesis

    Source        SS       DF        MS         F         P

Hypothesis   204.000        1   204.000    26.609     0.002
     Error    46.000        6     7.667
```

The last package design is significantly different from the other three taken as a group. Notice that the A matrix looks much different this time. Since the effects sum to zero, the last effect is minus the sum of the other three. That is, since

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$$

where α_i is the effect for level i of PACKAGE, then

$$\alpha_4 = -(\alpha_1 + \alpha_2 + \alpha_3)$$

and our contrast is

$$\alpha_1 + \alpha_2 + \alpha_3 - 3\alpha_4$$

which is

$$\alpha_1 + \alpha_2 + \alpha_3 - 3(-\alpha_1 - \alpha_2 - \alpha_3)$$

which simplifies to

$$4\alpha_1 + 4\alpha_2 + 4\alpha_3$$

If you don't like all this algebra, skip it. Remember, the **Contrast...** item does all this work automatically.

© 1989, SYSTAT, Inc.

Printing effects (beta coefficients)

If you choose [Show extended results] from **Data/Formats...**, SYSTAT displays the estimates of ANOVA effects (beta coefficients) in your model.

Multiple comparisons

You can do multiple comparisons in **MGLH** by using **Test...** several times, e.g.:

- Select **MGLH/Test...** from the **Stats** menu
- Specify PACKAGE for [Effects]
- Click [Contrast] and enter the matrix:
  1 -1 0 0
- Click OK

- Select **MGLH/Test...** from the **Stats** menu
- Specify PACKAGE for [Effects]
- Click [Contrast] and enter the matrix:
  1 0 -1 0
- Click OK

- Select **MGLH/Test...** from the **Stats** menu
- Specify PACKAGE for [Effects]
- Click [Contrast] and enter the matrix:
  1 0 0 -1
- Click OK

Unlike most multiple comparisons methods, post-hoc contrasts in **MGLH** are valid even when the cell counts are unequal.

You may wonder why you cannot stack all these contrasts together in a single operation. The programming to enable this is easy, but we did not do it for a reason. By making it more difficult to do a slew of uncritical, automatic comparisons, I am trying to help preserve the meaning of probabilities. *P*-values are a health hazard. The more of them you see on a computer output, the less meaningful they are.

© 1989, SYSTAT, Inc.

How do you control the overall comparison error rate for multiple hypotheses? The simplest approach is to use the Bonferroni procedure. If you want your overall rate to be α, then do each of k comparisons at α/k. That is, if you want the overall protection level for 5 comparisons to be .05, use multiple **Test...** runs with a critical level of .01 instead of .05. Remember, the Bonferroni method does not require a significant F-statistic for the factor you are doing the comparisons within. You can go straight to the Bonferroni comparisons without paying attention to the overall *F* value.

This method is extremely powerful, especially if you test a relatively small number of hypotheses. Furthermore, you are guaranteed not to exceed the nominal Type I error level (unlike some other multiple comparison procedures such as Duncan's multiple range test), although you may be conservative in rejecting the null hypothesis with a large number of comparisons. The multiple comparisons literature is enormous (see Miller, 1985 for a review), but a few points are worth noting:

1) You are wasting time if you look at the overall *F*-statistic before using multiple comparisons procedures like Tukey's HSD test (see the **Stats** chapter). These simultaneous test procedures are designed to maintain the overall protection level themselves. Oddly, many published studies show ANOVA tables and then use multiple comparison procedures on significant factors only. This is unnecessary.

2) All possible pairwise comparisons is a waste of power. Think about a meaningful subset of comparisons and **Test...** this subset with Bonferroni levels. You will almost always have more power than any other pairwise multiple comparisons procedure.

3) If you have an example where the Bonferroni method results in a different conclusion from some other pairwise comparison method, you are on slippery ground. The results probably won't replicate.

4) If you do an analysis of variance, look for a significant *F* value, and then confine the number of post-hoc comparisons to the degrees-of-freedom for the effect. Monte Carlo studies have shown that the nominal alpha level of the separate tests are comparable to that for the overall ANOVA (e.g. Carmer and Swanson, 1968; Bernhardson, 1975).

© 1989, SYSTAT, Inc.

5) If you are a statistician, choose [Show extended results] before using **Test...** SYSTAT prints information to help you compute Scheffé and other types of comparisons.

6) If journal editors or reviewers are still asking you to do Duncan's multiple-range test, please tell them that it is invalid. Question authority.

6.12 Two-way analysis of variance

Consider the following two-way analysis of variance design from Afifi and Azen (1972), cited in Kutner (1974) and reprinted in the BMDP and SAS manuals. The dependent variable, SYSINCR, is the change in systolic blood pressure after administering one of 4 different drugs to patients with one of 3 different diseases. Patients were assigned randomly to one of the 4 possible drugs.

	DISEASE 1	DISEASE 2	DISEASE 3
DRUG 1	42	33	31
	44	26	-3
	36	33	25
	13	21	25
	19		24
	22		
DRUG 2	28	34	3
	23	33	26
	34	31	28
	42	36	32
	13		4
			16
DRUG 3	1	11	21
	29	9	1
	19	7	9
		1	3
		-6	
DRUG 4	24	27	22
	9	12	7
	22	12	25
	-2	-5	5
	15	16	12
		15	

© 1989, SYSTAT, Inc.

The data are set up in the SYSTAT file AFIFI in the following way:

DRUG	DISEASE	SYSINCR
1	1	42
1	1	44
1	1	36
.	.	.
.	*etc.*	.
.	.	.
4	3	5
4	3	12

Now we can perform a least squares two-way analysis of variance.

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variable: SYSINCR
- Click [Factors] and select the factors, specifying the number of levels for each: DRUG (4 levels), DISEASE (3)
- Click [Save residuals]; click [Save residuals and diagnostics]
- Click OK

You must specify the number of levels for each factor. Because this is a factorial design, **Fully factorial (M)ANOVA...** automatically generates an interaction term (DRUG*DISEASE). If you like pointing and clicking, you could do the same analysis with **General linear model...**:

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: SYSINCR
- Click [Indep'dent] and select the independent variables: DRUG, DISEASE
- Click [Interactions] and select the interaction DRUG*DISEASE
- Click [Categories] and select the categorical variables, specifying the number of levels for each variable: DRUG (4 levels), and DISEASE (3 levels)
- Click [Save file] and choose [Save residuals and diagnostics]
- Click OK

© 1989, SYSTAT, Inc.

Both methods of setting up the model produce the same output:

```
Dep Var: SYSINCR      N:  58   Multiple R:  .675   Squared Multiple R:  .456

                    Analysis of Variance

Source          Sum-of-Squares   DF  Mean-Square    F-Ratio      P

DRUG               2997.472       3     999.157       9.046     0.000
DISEASE             415.873       2     207.937       1.883     0.164
DRUG*
DISEASE             707.266       6     117.878       1.067     0.396

Error              5080.817      46     110.453
```

If you have an orthogonal design (equal number of cases in every cell), you can stop here. The ANOVA table is the same one you get with any standard ANOVA program. **MGLH** can handle non-orthogonal designs, however, as in the present example. To understand the sources for sums of squares, you must know something about least squares ANOVA.

As with one-way ANOVA, your specifying factor levels causes **MGLH** to create dummy variables out of the classifying input variable. SYSTAT creates one fewer dummy variables than categories specified. Factors must have integer values from 1 to *k*. Any value outside that range causes an error (use DATA to recode a variable as needed for **MGLH**). If any cases have a value of 0, **MGLH** codes all the *k*–1 dummy variables for that observation as 0. This is not like having missing data. It is a device that can be used for creating nested designs. More about that later in Example 6.19, "Nested designs."

Coding of the dummy variables is the classic analysis of variance parameterization, in which the sum of effects estimated for a classifying variable is zero (Scheffé, 1959). In our example, DRUG has 4 categories. The [Categories] option therefore creates 3 dummy variables with the following scores for observations at each level:

1 0 0	for observations under DRUG=1
0 1 0	for observations under DRUG=2
0 0 1	for observations under DRUG=3
–1 –1 –1	for observations under DRUG=4

© 1989, SYSTAT, Inc.

Since DISEASE has 3 categories, the [Categories] option causes 2 dummy variables to be coded as follows:

1 0	for observations under DISEASE=1
0 1	for observations under DISEASE=2
–1 –1	for observations under DISEASE=3

Now, since there are no continuous predictors in the model (unlike the analysis of covariance), we have a complete design matrix of dummy variables as follows (DRUG is labeled with an A, DISEASE with a B, and the grand mean with an m):

TREATMENT		*DESIGN MATRIX*											
		mean	*drug effects*			*disease effects*		*interaction effects*					
A	B	m	a1	a2	a3	b1	b2	a1b1	a1b2	a2b1	a2b2	a3b1	a3b2
1	1	1	1	0	0	1	0	1	0	0	0	0	0
1	2	1	1	0	0	0	1	0	1	0	0	0	0
1	3	1	1	0	0	–1	–1	–1	–1	0	0	0	0
2	1	1	0	1	0	1	0	0	0	1	0	0	0
2	2	1	0	1	0	0	1	0	0	0	1	0	0
2	3	1	0	1	0	–1	–1	0	0	–1	–1	0	0
3	1	1	0	0	1	1	0	0	0	0	0	1	0
3	2	1	0	0	1	0	1	0	0	0	0	0	1
3	3	1	0	0	1	–1	–1	0	0	0	0	–1	–1
4	1	1	–1	–1	–1	1	0	–1	0	–1	0	–1	0
4	2	1	–1	–1	–1	0	1	0	–1	0	–1	0	–1
4	3	1	–1	–1	–1	–1	–1	1	1	1	1	1	1

If you use **General linear model...** instead of **Fully factorial (M)ANOVA...**, you must select the variables in the order above if you want the columns to correspond to this example.

© 1989, SYSTAT, Inc.

The reason for going through all of this is to explain how MGLH gets an error term for the ANOVA table. Since it is a least squares program, the error term is taken from the residual sum of squares in the regression onto the above dummy variables. For nonorthogonal designs, this choice is identical to that produced by BMDP2V and SAS PROC GLM with Type III sum of squares. These, in general, will be the hypotheses you want to test on unbalanced experimental data. You can construct other types of sums of squares (such as those in SAS GLM) by using [A matrix] in **Test...**, but you should consult the references if you do not already know how to do so. An example is given below in the section on unbalanced designs.

Post-hoc tests

It is evident that only the main effect for DRUG is significant. Therefore, we might want to test some contrasts on the DRUG effects. We could do any contrast by reading in an A matrix, but a simpler way is to use [Contrast] to enter one. For example, to contrast the first and third drugs, do this:

- Select **MGLH/Test...** from the **Stats** menu
- Specify DRUG for [Effects]
- Click [Contrast] and enter the matrix:
  1 0 –1 0
- Click OK

You need four numbers (1 0 –1 0) because there are four levels of DRUG. Notice the A matrix created from this input in the output below. [Contrast] automatically takes into account the degree of freedom lost in the dummy coding. Also note that you do not need to normalize contrasts or rows of the A matrix to unit vector length, as in some ANOVA programs. If you use (2 0 –2 0) or (.707 0 –.707 0) instead of (1 0 –1 0), you get the same sum of squares. It should be evident that you can use [Contrast] only if you used the [Categories] option of **General linear model...** or the [Factors] option of **Fully factorial (M)ANOVA...**. Here is the result the above hypothesis test.

© 1989, SYSTAT, Inc.


```
Test for effect called:
                DRUG

A Matrix

          1         2         3         4         5
      0.000     1.000     0.000    -1.000     0.000
          6         7         8         9        10
      0.000     0.000     0.000     0.000     0.000
         11        12
      0.000     0.000

Test of Hypothesis

    Source        SS       DF        MS         F         P

Hypothesis  1697.545        1  1697.545    15.369     0.000
     Error  5080.817       46   110.453
```

You may not use [Contrast] on an interaction term. It creates an A matrix only for main effects.

Here is one more example for testing the first against the fourth drug:

- Select **MGLH/Test...** from the **Stats** menu
- Specify DRUG for [Effects]
- Click [Contrast] and enter the matrix:
  1 0 0 –1
- Click OK

```
Test for effect called:
                DRUG

A Matrix

          1         2         3         4         5
      0.000     2.000     1.000     1.000     0.000
          6         7         8         9        10
      0.000     0.000     0.000     0.000     0.000
         11        12
      0.000     0.000

Test of Hypothesis

    Source        SS       DF        MS         F         P

Hypothesis  1178.892        1  1178.892    10.673     0.002
     Error  5080.817       46   110.453
```

© 1989, SYSTAT, Inc.

Notice that in the A matrix created by this contrast, we get (2 1 1) in the place of the three dummy variables corresponding to the appropriate columns of the A matrix. If you do not see why this happens, read on...

Because we selected the reduced form for coding of dummy variables in which sums of effects are zero, we have the following restriction for the DRUG effects:

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$$

where α_i is the effect for that level of DRUG. This means that

$$\alpha_4 = -(\alpha_1 + \alpha_2 + \alpha_3)$$

and the contrast DRUG(1)–DRUG(4) is equivalent to:

$$\alpha_1 - [-(\alpha_1 + \alpha_2 + \alpha_3)] = 0$$

which is

$$2\alpha_1 + \alpha_2 + \alpha_3$$

Tests of simple effects

Now we can do simple contrasts between drugs within levels of disease, although the lack of a significant DRUG*DISEASE interaction does not justify it. To show how it is done, we consider a contrast between the first and third levels of DRUG for the first DISEASE only. We must specify the contrast in terms of the cell means. Using the terminology

$$\text{MEAN}(\textit{DRUG index}, \textit{DISEASE index}) = \mu_{ij}$$

we want to contrast cell means μ_{12} and μ_{31}. These are composed of:

$$\mu_{11} = \mu + \alpha_1 + \beta_1 + \alpha\beta_{11}$$

$$\mu_{31} = \mu + \alpha_3 + \beta_1 + \alpha\beta_{31}$$

The difference between these two means is therefore

$$\mu_{11} - \mu_{31} = \alpha_1 - \alpha_3 + \alpha\beta_{11} - \alpha\beta_{31}$$

© 1989, SYSTAT, Inc.

Now, if you look back at the illustration of the design matrix above, you can find these terms and construct an A matrix that picks up each one at the appropriate column. Here are the column labels of the design matrix ("a" means DRUG and "b" means DISEASE) to serve as a [column ruler] over the A matrix specified in this hypothesis paragraph. (You can use a single space between columns—we have extra space here to accommodate the labels in the "column ruler.")

DESIGN MATRIX

m	a1	a2	a3	b1	b2	a1b1	a1b2	a2b1	a2b2	a3b1	a3b2

- Select **Test...** from the **Stats** menu
- Click [A matrix] and enter the matrix, using spaces between columns:

  0 1 0 −1 0 0 1 0 0 0 −1 0
- Click OK

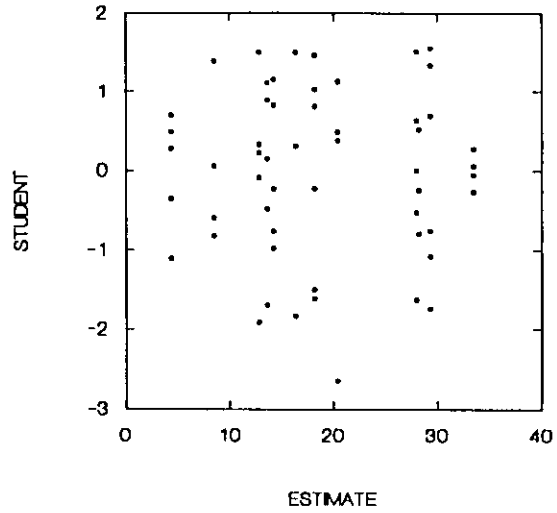
Once you understand how your specifying factor levels (or, with the command interface, the CATEGORY sentence) codes design variables and how the model sentence orders them, you can take any standard ANOVA text like Winer (1971) or Scheffé (1959) and construct an A matrix for any linear contrast.

Incidentally, if you want cell means and other within-cell statistics, use **Stats** and compute basic statistics **By groups...** Also, if you [Save residuals], the values of the ESTIMATE variable are the least squares cell means.

Screening results: ANOVA is not robust against everything

Let's examine these data in more detail. We clicked [Save residuals] to save the residuals in a file. Here is how we might analyze them to examine our ANOVA assumptions. First, we plot residuals against estimated values (cell means) to check for homogeneity of variance. We use the Studentized residuals because it is convenient to reference them against a *t* distribution. Note the value at the bottom of the following plot badly skews the data in its cell (which happens to be DRUG1,DISEASE3).

© 1989, SYSTAT, Inc.



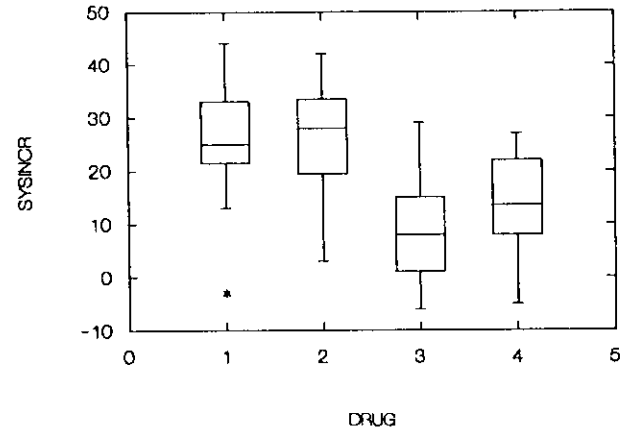
Here is a stem-and-leaf plot of the studentized residuals.

```
        Stem and Leaf Plot of Variable:  STUDENT    , N =   58

Minimum is:        -2.647
Lower Hinge is:      -0.761
Median is:          0.101
Upper Hinge is:       0.698
Maximum is:         1.552

              -2   6
              -2
              -1   987666
              -1   410
              -0 H 9877765
              -0   432222000
               0 M 0001222333444
               0 H 55666888
               1   011133444
               1   55
```

Notice that the smallest value seems to be out of line. A *t* statistic value of –2.647 corresponds to $p<.01$, and we would not expect one this small to show up in a sample of only 58 independent values. This is the same point that appeared at the bottom of the previous plot.

Now look at the box plot of the dependent variable against the four drug groups. The outlier stands out clearly in the first group.

© 1989, SYSTAT, Inc.



To see the effect of this outlier, we used DATA to DELETE the observation with the small studentized residual. Then, we ran the analysis again. Here is the **MGLH** output on the revised data:

```
Dep Var: SYSINCR      N:  57    Multiple R:  .710    Squared Multiple R:  .503

                        Analysis of Variance

  Source     Sum-of-Squares    DF  Mean-Square     F-Ratio       P

    DRUG          3344.064     3     1114.688      11.410     0.000
 DISEASE           232.826     2      116.413       1.192     0.313

    DRUG*
 DISEASE           676.865     6      112.811       1.155     0.347

   Error          4396.367    45       97.697
```

The differences are not earth-shattering. Nevertheless, notice that the disease effect is substantially attenuated when only one case out of 58 is deleted. Daniel (1960) gives an example in which one outlying case alters the fundamental conclusions of a designed experiment. The F-test is robust to certain violations of assumptions, but factorial ANOVA is not robust against outliers. You should routinely do these plots for ANOVA.

© 1989, SYSTAT, Inc.

6.13 Single degree of freedom designs

The following data (REACT) involve yields of a chemical reaction under various combinations of four binary factors (A,B,C,D). Two reactions were observed under each combination of experimental factors, so the number of cases per cell is 2.

A	B	C	D	YIELD
1	1	1	1	15
1	1	1	1	779
1	1	1	2	999
1	1	1	2	990
1	1	2	1	499
1	1	2	1	212
1	1	2	2	286
1	1	2	2	611
1	2	1	1	438
1	2	1	1	239
1	2	1	2	926
1	2	1	2	787
1	2	2	1	871
1	2	2	1	303
1	2	2	2	891
1	2	2	2	663
2	1	1	1	177
2	1	1	1	831
2	1	1	2	222
2	1	1	2	182
2	1	2	1	342
2	1	2	1	426
2	1	2	2	416
2	1	2	2	910
2	2	1	1	474
2	2	1	1	328
2	2	1	2	70
2	2	1	2	121
2	2	2	1	569
2	2	2	1	111
2	2	2	2	771
2	2	2	2	119

Let's analyze these data in a four-way ANOVA:

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variable: YIELD
- Click [Factors] and select the factors, specifying the number of levels for each: A (2 levels), B (2 levels), C (2 levels), D (2 levels)
- Click OK

© 1989, SYSTAT, Inc.

You can see the advantage of **Fully factorial (M)ANOVA...** over **General linear model...** when you have lots of factors: you only have to select the independent variables (factors) once, and you specify the levels right away. With **General linear model...**, you would have to select the factors as [Indep'dent] variables first and then select all the [Interactions] and finally select all the [Categorical] variables all over again, this time specifying the number of levels.

```
Dep Var:   YIELD      N:   32    Multiple R:  .755    Squared Multiple R:  .570

                          Analysis of Variance

Source      Sum-of-Squares    DF  Mean-Square     F-Ratio       P

A            369800.000        1  369800.000        4.651      0.047
B              1458.000        1    1458.000        0.018      0.894
C              5565.125        1    5565.125        0.070      0.795
D            172578.125        1  172578.125        2.170      0.160
A*
B             87153.125        1   87153.125        1.096      0.311
A*
C            137288.000        1  137288.000        1.727      0.207
A*
D            328860.500        1  328860.500        4.136      0.059
B*
C             61952.000        1   61952.000        0.779      0.390
B*
D              3200.000        1    3200.000        0.040      0.844
C*
D              3160.125        1    3160.125        0.040      0.844
A*
B*
C             81810.125        1   81810.125        1.029      0.325

A*
B*
D              4753.125        1    4753.125        0.060      0.810
A*
C*
D            415872.000        1  415872.000        5.230      0.036
B*
C*
D              4.500           1       4.500        0.000      0.994
A*
B*
C*
D             15051.125        1   15051.125        0.189      0.669

Error       1272247.000       16   79515.438
```

We have a significant main effect for the first factor (A) plus one significant interaction (A*C*D). Let's look at the study more closely.

© 1989, SYSTAT, Inc.

Since this is a single degree of freedom study (a 2^n factorial), if the usual assumptions for the experiment are valid, each effect estimate is normally distributed. All the effects estimates except the constant have zero mean and common variance (since dummy variables were used in their computation). Thus, we can compare them to a normal distribution. Let's save the effects into a SYSTAT file. Since SYSTAT remembers our last selections, we can just select **MGLH/Fully factorial (M)ANOVA...**, click [Save residuals], and click OK. The model is still set up.

- Select **MGLH**/Fully factorial (M)ANOVA... from the Stats menu
- Click [Save residuals]
- Choose [Save regression coefficients]
- Click OK

This reestimates the model and saves the regression coefficients (effects). The file has one "case" with 16 variables (CONSTANT plus 15 effects). The effects are labeled X(1), X(2), and so on, because they are related to the dummy variables, not the original variables A,B,C, and D. Let's transpose this file into a new file containing only the 15 effects.

- Select **Transpose...** from the **Data** menu
- Specify a filename for the new file
- Click OK

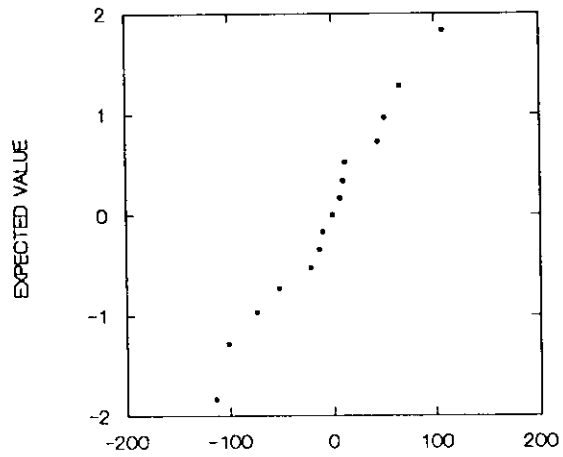
- Select **Open...** from the **File** menu
- Select the transposed file
- Click OK

- Edit the new file and delete the value for CONSTANT
- Click OK

Now we can open the new file and do a probability plot of the effects. The variable we need is COL(1).

- Select **Prob'y/Normal...** from the **Graph** menu
- Click OK

© 1989, SYSTAT, Inc.



These effects are indistinguishable from a random normal variable. They plot almost on a straight line. What does it mean for the study? For the "significant" *F* tests?

It's time for us to tell you that the data were produced by a random number generator. Obviously we had some lessons in mind.

1) If you are doing a factorial analysis of variance, the *p* values you see on the output are not adjusted for the number of factors. If you do a 3 way design, you look at 7 tests (excluding the constant). For a 4 way design, you examine 15 tests. For 5 or more factors, you are insane. Out of 15 *F* tests on random data, expect on average to find at least one test approaching significance. We got 2 significant and 1 almost significant, which is not far out of line. We have tried to find a discussion of this issue in popular ANOVA texts. No one seems to point out that the probabilities for each separate *F* test need to be corrected for the experiment-wise error rate. Some authors devote chapters to fine distinctions between multiple comparison procedures and then illustrate them within a multi-factorial design uncorrected for the experiment-wise error rate we have just demonstrated. Remember this: a factorial design *is* a multiple comparison.

© 1989, SYSTAT, Inc.

2) If you have a single degree of freedom study, use the procedure we used to draw a probability plot of the effects. Any effect that is really significant will stand out like a sore thumb.

3) If you have a factorial study with more degrees of freedom on some factors, use the Bonferroni critical value for deciding which effects are significant. It guarantees that the Type I error rate for the study will be no greater than the level you choose. In the above example, this value is .05/15, i.e. .003. We would not have been fooled.

4) Multiple *F* tests based on a common denominator (mean square error in this example) are correlated. This complicates the problem further. In general, the greater the discrepancy between numerator and denominator degrees of freedom and the smaller the denominator degrees of freedom, the greater the dependence of the tests. The Bonferroni tests are your best defense in this situation, though Feingold and Korsog (1986) offer some useful alternatives.

6.14 Unbalanced designs

An unbalanced factorial design occurs when the numbers of cases in cells are unequal and not proportional across rows or columns. Here is an example of a two-by-two design:

	B1	B2
A1	1 2	5 3 4
A2	6 7 9 8 4	2 1 5 3

Unbalanced designs require a least squares ANOVA program like **MGLH** because the usual maximum likelihood method of adding up sums of squared deviations from cell means and the grand mean does not yield maximum likelihood estimates of effects. **MGLH** adjusts for unbalanced designs when you get an ANOVA table or use the [Effects] option in testing hypotheses (**Test...**). A few other programs (e.g. SAS GLM, BMDP4V, SPSS MANOVA) make these same adjustments. Programs that use "unweighted means" or "harmonic means" do not.

© 1989, SYSTAT, Inc.

You are not necessarily home free with a least squares program, however. The estimates of effects in the unbalanced design are no longer orthogonal (and thus statistically independent) across factors and their interactions. This means that the sum of squares associated with one factor depends on the sum of squares for another or its interaction.

Analysts accustomed to using multiple regression have no problem with this situation because they assume that their independent variables in a model are correlated. Experimentalists, however, often have difficulty speaking of a main effect "conditioned" on another. Consequently, there is a large literature on hypothesis testing methodology for unbalanced designs (e.g. Speed and Hocking, 1976, and Speed, Hocking, and Hackney,1978), and there is no consensus on how to test hypotheses with nonorthogonal designs.

Some statisticians advise you to do a series of hierarchical tests beginning with interactions. If the highest order interactions are insignificant, drop them from the model and recompute the analysis. Then examine the lower order interactions. If they are insignificant, recompute the model with main effects only. And so on. Some computer programs automate this process and print sums of squares and *F* tests according to the hierarchy (ordering of effects) you specify in the model. SAS GLM, for example, calls these Type I sums of squares.

This procedure is analogous to stepwise regression, in which hierarchical subsets of models are tested. Here is an example. Assume you have specified the following model:

Y = CONSTANT + A + B + C + A*B + A*C + B*C + A*B*C

The hierarchical approach tests the following models:

Y = CONSTANT + A + B + C + A*B + A*C + B*C + A*B*C
Y = CONSTANT + A + B + C + A*B + A*C + B*C
Y = CONSTANT + A + B + C + A*B + A*C
Y = CONSTANT + A + B + C + A*B
Y = CONSTANT + A + B + C
Y = CONSTANT + A + B
Y = CONSTANT + A

© 1989, SYSTAT, Inc.

The problem with this approach, however, is that plausible subsets of effects are ignored if you examine only one hierarchy. The following model, which may be the best fit to the data, is never considered:

Y = CONSTANT + A + B + A*B

Furthermore, if you do decide to examine all the other plausible subsets, you are really doing all possible subsets regression and you had better use Bonferroni confidence levels before rejecting a null hypothesis. Our example above has 127 possible subset models (excluding ones without a CONSTANT). That's a lot of models to snoop among.

You could leave out models that do not contain all lower order effects within each higher order effect. Eliminate models like the following, for example:

Y = CONSTANT + A + A*B*C

If you do this, you are left with 15 models to consider in our example. You still cannot consider the probabilities on the output as real (at the nominal error rate) unless you modify them via Bonferroni.

At this point, you may have deduced our opinion. We believe that if you have done an experiment and have decided that higher order effects (interactions) are of enough theoretical importance to include in your model, then you are stuck with them. In other words, you should condition every test on all other effects in the model you selected. This is the classical approach of Fisher and Yates, for example. It amounts to using the default *F* values on the **MGLH** output, which are the same as the SAS Type III sums of squares. (For historical reasons, SAS called the sums of squares that are used most frequently "Type III" and those used least frequently "Type I.")

Probably the most important reason to follow our strategy and stay with one model is that if you eliminate a series of effects that are not quite significant (say, p=.06) you could end up with a totally incorrect subset model because of the dependencies among the sums of squares. In summary, if you want other sums of squares in **MGLH**, go ahead and compute them. You can use the [Error] option in **Test...** to customize sums of squares. We refuse, however, to print all sorts of sums of squares automatically and then pretend that the p values mean something.

© 1989, SYSTAT, Inc.

6.15 Randomized block designs

A randomized block design is like a factorial design without an interaction term. Here is an example from Neter, Wasserman, and Kutner (1985). Five blocks of judges were given the task of analyzing three treatments. Subjects (judges) are stratified within blocks, so the interaction of blocks and treatments cannot be analyzed. These data are in the file BLOCK.

BLOCK	TREAT	JUDGMENT
1	1	1
1	2	5
1	3	8
2	1	2
2	2	8
2	3	14
3	1	7
3	2	9
3	3	16
4	1	6
4	2	13
4	3	18
5	1	12
5	2	14
5	3	17

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: JUDGMENT
- Click [Indep'dent] and select the independent variables: BLOCK, TREAT
- Click [Categories] and select the categorical variables, specifying the number of levels for each: BLOCK (5 levels), TREAT (3 levels)
- Click OK

We must use **General linear model...** instead of **Fully factorial (M)ANOVA...** because we do not want the BLOCK*TREAT interaction in the model.

```
Dep Var:JUDGMENT      N:   15    Multiple R:  .970    Squared Multiple R:  .940

                        Analysis of Variance

Source      Sum-of-Squares    DF   Mean-Square      F-Ratio        P

BLOCK          171.333         4        42.833        14.358      0.001
TREAT          202.800         2       101.400        33.989      0.000

Error           23.867         8         2.983
```

© 1989, SYSTAT, Inc.

6.16 Incomplete block designs

Randomized blocks can be used in factorial designs as well. Here is an example from John (1971). The data (in the file JOHN) involve an experiment with 3 treatment factors (A, B, and C) plus a blocking variable with 8 levels. Note that data were collected on 32 of the possible 64 experimental situations.

BLOCK	A	B	C	Y
1	1	1	1	101
1	2	1	2	373
1	1	2	2	398
1	2	2	1	291
2	1	1	2	312
2	2	1	1	106
2	1	2	1	265
2	2	2	2	450
3	1	1	1	106
3	2	2	1	306
3	1	1	2	324
3	2	2	2	449
4	1	2	1	272
4	2	1	1	89
4	1	2	2	407
4	2	1	2	338
5	1	1	1	87
5	2	1	2	324
5	1	2	1	279
5	2	2	2	471
6	1	1	2	323
6	2	1	1	128
6	1	2	2	423
6	2	2	1	334
7	1	1	1	131
7	2	1	1	103
7	1	2	2	445
7	2	2	2	437
8	1	1	2	324
8	2	1	2	361
8	1	2	1	302
8	2	2	1	272

© 1989, SYSTAT, Inc.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: Y
- Click [Indep'dent] and select the independent variables: BLOCK, A, B, C
- Click [Interactions] and select the interactions, clicking [Next] between interactions: A*B, A*C, A*B*C
- Click [Categories] and select the categorical variables, specifying the number of levels for each : BLOCK (8 levels), A (2 levels), B (2 levels), C (2 levels)
- Click OK

Again, we must use **General linear model...** instead of **Fully factorial (M)ANOVA...** because we do not want interactions with BLOCK in the model.

```
Dep Var:      Y     N:   32    Multiple R:  .994    Squared Multiple R:  .988

                          Analysis of Variance

Source      Sum-of-Squares    DF  Mean-Square     F-Ratio        P

BLOCK          2638.469        7      376.924        1.182      0.364
A              3465.281        1     3465.281       10.862      0.004
B            161170.031        1   161170.031      505.209      0.000
C            278817.781        1   278817.781      873.992      0.000
A*
B              28.167          1       28.167        0.088      0.770
A*
C            1802.667          1     1802.667        5.651      0.029
B*
C            11528.167         1    11528.167       36.137      0.000
A*
B*
C              45.375          1       45.375        0.142      0.711

Error        5423.281         17      319.017
```

6.17 Fractional factorial designs

Sometimes a factorial design involves so many combinations of treatments that certain cells must be left empty to save experimental resources. Other times, a complete randomized factorial study is designed, but loss of subjects leaves one or more cells completely missing. These models are similar to incomplete block designs in that not all effects in the full model can be estimated. Usually, certain interactions must be left out of the model.

© 1989, SYSTAT, Inc.

Here is an example on some experimental data that contain values in only 8 out of 16 possible cells. Each cell contains 2 cases. The pattern of nonmissing cells makes it possible to estimate only the main effects plus 3 two-way interactions. The data are in the file FRACTION.

A	B	C	D	Y
1	1	1	1	7
1	1	1	1	3
2	2	1	1	1
2	2	1	1	2
2	1	2	1	12
2	1	2	1	13
1	2	2	1	14
1	2	2	1	15
2	1	1	2	8
2	1	1	2	6
1	2	1	2	12
1	2	1	2	10
1	1	2	2	6
1	1	2	2	4
2	2	2	2	6
2	2	2	2	7

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: Y
- Click [Indep'dent] and select the independent variables: A, B, C, D
- Click [Interactions] and select the interactions, clicking [Next] between interactions: A*B, A*C, B*C
- Click [Categories] and select the categorical variables, specifying the number of levels for each variable: A (2 levels), B (2 levels), C (2 levels), D (2 levels)
- Click OK

Again, we must use **General linear model...** instead of **Fully factorial (M)ANOVA...** to omit the higher way interactions that **Fully factorial (M)ANOVA...** automatically generates.

© 1989, SYSTAT, Inc.


```
Dep Var:      Y     N:   16    Multiple R:  .972    Squared Multiple R:  .944

                        Analysis of Variance

Source      Sum-of-Squares    DF  Mean-Square    F-Ratio       P

A                   16.000     1       16.000      8.000   0.022
B                    4.000     1        4.000      2.000   0.195
C                   49.000     1       49.000     24.500   0.001
D                    4.000     1        4.000      2.000   0.195
A*
B                  182.250     1      182.250     91.125   0.000
A*
C                   12.250     1       12.250      6.125   0.038
B*
C                    2.250     1        2.250      1.125   0.320

Error               16.000     8        2.000
```

When missing cells turn up by happenstance rather than by design, you may not know which interactions to eliminate. When you attempt to fit the full model, **MGLH** informs you that the design is singular. In that case, you may need to try several models before finding an estimable one. It is usually best to begin by leaving out the highest order interaction (A*B*C*D in this example). Select **General linear model...** again, click OK, and see what happens. Continue with subset models until you get an ANOVA table.

Looking for an estimable model is not the same as data snooping with stepwise regression because you are not looking at *p* values. Once you find an estimable model, you stop and settle with the statistics printed in the ANOVA table.

6.18 Mixed models

Mixed models involve combinations of fixed and random factors in an ANOVA. Fixed factors are assumed to be comprised of an exhaustive set of categories (e.g., males and females), while random factors have categories levels that are assumed to have been randomly sampled from a larger population of categories (e.g., classrooms or word stems). Because of the mixing of fixed and random components, expected mean squares for certain effects are different from those for fully fixed or fully random designs. **MGLH** can handle mixed models because **Test...** has an [Error] option to select error terms for specific hypotheses.

© 1989, SYSTAT, Inc.

For example, let's reanalyze the AFIFI dataset with a mixed model instead of a fully fixed factorial. This time, we are interested in the 4 drugs as wide-spectrum disease killers. Since each drug is now thought to be a shotgun against diseases in general, we have sampled 3 random diseases to assess the drugs. This implies that DISEASE is a random factor and DRUG remains a fixed factor. In this case, the error term for DRUG works out to the DRUG*DISEASE interaction. Here is how to test the effect. Run the same analysis we already performed to get the ANOVA table:

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variable: SYSINCR
- Click [Factors] and select the factors, specifying the number of levels for each: DRUG (4 levels), DISEASE (3 levels)
- Click OK

The output for this stage is shown in a previous example. Now specify DRUG*DISEASE as the [Error] term for the DRUG effect in **Test...**:

- Select **MGLH/Test...** from the **Stats** menu
- Specify DRUG for [Effects]
- Specify DRUG*DISEASE for [Error], using an asterisk to indicate interaction
- Click OK

```
Test for effect called:
                DRUG

Test of Hypothesis

     Source        SS       DF        MS          F         P

Hypothesis    2997.472        3   999.157      8.476     0.014
     Error     707.266        6   117.878
```

Of course, you could divide the mean squares yourself from the ANOVA table above to get the *F* value (8.476). For MANOVA mixed models (see Example 6.25, "Multivariate analysis of variance"), that is more difficult. The [Error] option is handy.

© 1989, SYSTAT, Inc.

6.19 Nested designs

Nested designs resemble factorial designs with certain cells missing (incomplete factorials). This is because one factor is "nested" under another so that not all combinations of the two factors are observed. For example, in an educational study, classrooms are usually nested under schools because it is impossible to have the same classroom existing at two different schools (except as antimatter). The following example is from Neter, Wasserman, and Kutner (1985), in which teachers are nested within schools. The data (learning scores) look like this:

	TEACHER1	TEACHER2
SCHOOL1	25 29	14 11
SCHOOL2	11 6	22 18
SCHOOL3	17 20	5 2

In the study there are actually 6 teachers, not just 2. Thus, the design really looks like this:

	TEACHER1	TEACHER2	TEACHER3	TEACHER4	TEACHER5	TEACHER6
SCHOOL1	25 29	14 11				
SCHOOL2			11 6	22 18		
SCHOOL3					17 20	5 2

If we allow TEACHER to have 6 levels, however, we are not letting the experimental design follow the actual design of the study. We want TEACHER to have 2 levels, but we need an indicator of each teacher's school. To do this, we use three TEACHER variables, one for each school. In the input data below, read TEACHER(1) as "teacher within school 1," and so on. Notice that a value of 0 on a TEACHER variable means that the teacher was not in the respective school.

© 1989, SYSTAT, Inc.

The data are set up in the file SCHOOLS.

TEACHER(1)	TEACHER(2)	TEACHER(3)	SCHOOL	LEARNING
1	0	0	1	25
1	0	0	1	29
2	0	0	1	14
2	0	0	1	11
0	1	0	2	11
0	1	0	2	6
0	2	0	2	22
0	2	0	2	18
0	0	1	3	17
0	0	1	3	20
0	0	2	3	5
0	0	2	3	2

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: LEARNING
- Click [Indep'dent] and select the independent variables: TEACHER(1), TEACHER(2), TEACHER(3), SCHOOL
- Click [Categories] and select the categorical variables, specifying the number of levels for each: TEACHER(1) (2 levels), TEACHER(2) (2 levels), TEACHER(3) (2 levels), and SCHOOL (3 levels)
- Click OK

```
Dep Var:LEARNING      N:   12    Multiple R:  .972    Squared Multiple R:  .945

                          Analysis of Variance

  Source    Sum-of-Squares    DF  Mean-Square     F-Ratio       P

TEACHER(1)        210.250     1      210.250      30.036      0.002
TEACHER(2)        132.250     1      132.250      18.893      0.005
TEACHER(3)        225.000     1      225.000      32.143      0.001
 SCHOOL           156.500     2       78.250      11.179      0.009

  Error            42.000     6        7.000
```

Notice that the TEACHER effect is split across 3 single degree-of-freedom effects. This decomposition is nice, but we want one 3 degree-of-freedom test of TEACHER. Remember how we used ampersand (&) to test groups of coefficients together? Here is how to do it:

- Select **MGLH/Test...** from the **Stats** menu
- Select [Effects]: TEACHER(1) & TEACHER(2) & TEACHER(3)
- Click OK

© 1989, SYSTAT, Inc.


```
Test for effect called:
                TEACHER(1)
          and
                TEACHER(2)
          and
                TEACHER(3)

A Matrix

                     1          2          3          4          5

          1      0.000      1.000      0.000      0.000      0.000
          2      0.000      0.000      1.000      0.000      0.000
          3      0.000      0.000      0.000      1.000      0.000

                     6

          1      0.000
          2      0.000
          3      0.000

Test of Hypothesis

     Source         SS        DF         MS          F          P

Hypothesis     567.500         3    189.167     27.024      0.001
     Error      42.000         6      7.000
```

Neter, Wasserman, and Kutner go on to describe how to do nested unbalanced designs (unequal number of cases per cell). The nice thing about **MGLH** is that unbalanced designs are handled automatically because the estimation method is least squares. In balanced cases, like the one above, the results agree with programs that can handle only balanced designs.

6.20 Split plot designs

The split plot design is closely related to the nested design. In the split plot, however, plots are often considered a random factor. Thus we have to construct different error terms to test different effects. Here is an example involving two treatments: A (between plots) and B (within plots). The numbers in the cells are YIELD of the crop within plots.

	A1		A2	
	PLOT1	PLOT2	PLOT3	PLOT4
B1	0	3	4	5
B2	0	1	2	4
B3	5	5	7	6
B4	3	4	8	6

To input the data, as in the nested design, we must use two variables to represent PLOT. The first PLOT(1) represents PLOT within A1. The second PLOT(2) represents PLOT within A2. When these variables have a value of 0, it means that the respective case is not in the respective plot.

© 1989, SYSTAT, Inc.

PLOT(1)	PLOT(2)	A	B	YIELD
1	0	1	1	0
1	0	1	2	0
1	0	1	3	5
1	0	1	4	3
2	0	1	1	3
2	0	1	2	1
2	0	1	3	5
2	0	1	4	4
0	1	2	1	4
0	1	2	2	2
0	1	2	3	7
0	1	2	4	8
0	2	2	1	5
0	2	2	2	4
0	2	2	3	6
0	2	2	4	6

To analyze this design, we need two different error terms. For the Between Plots effects (A), we need "PLOTS within A." For the Within Plots effects (B and A*B), we need "B by PLOTS within A."

Fit the saturated model

First, we fit the saturated model with all the effects and then we use **Test...** to pick up the different error terms.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: YIELD
- Click [Indep'dent] and select the independent variables: A, B, PLOT(1), PLOT(2)
- Click [Interactions] and select the interactions, clicking [Next] between each interaction: A*B, B*PLOT(1), B*PLOT(2)
- Click [Categories] and select the categorical variables, specifying the number of levels for each variable: PLOT(1) (2 levels), PLOT(2) (2 levels), A (2 levels), and B (4 levels)
- Click OK

```
Dep Var:   YIELD      N:   16    Multiple R: 1.000    Squared Multiple R: 1.000
```

Between plots effects

We get no ANOVA table because the model is perfectly fit. The coefficient of determination (Squared Multiple R) is 1. Now we have to use some of the effects as error terms. First, the test for between plots effects, namely A:

© 1989, SYSTAT, Inc.

- Select **MGLH/Test...** from the **Stats** menu
- Specify [Effects]: A
- Specify [Error], using an ampersand: PLOT(1) & PLOT(2)
- Click OK

```
Test for effect called:
                A

Test of Hypothesis

     Source        SS       DF        MS          F          P

Hypothesis     27.563        1     27.563     17.640      0.052
     Error      3.125        2      1.563
```

Within plots effects

Then we do the within plots effects. First B:

- Select **MGLH/Test...** from the **Stats** menu
- Specify [Effects]: B
- Specify [Error], using an ampersand and asterisks: B*PLOT(1) & B*PLOT(2)
- Click OK

```
Test for effect called:
                B

Test of Hypothesis

     Source        SS       DF        MS          F          P

Hypothesis     42.688        3     14.229     11.576      0.007
     Error      7.375        6      1.229
```

Finally, we do A*B with the same error term:

- Select **MGLH/Test...** from the **Stats** menu
- Specify [Effects], using an asterisk: A*B
- Specify [Error], using an ampersand and asterisks: B*PLOT(1) & B*PLOT(2)
- Click OK

```
Test for effect called:
                A
          BY
                B

Test of Hypothesis

     Source        SS       DF        MS          F          P

Hypothesis      2.188        3      0.729      0.593      0.642
     Error      7.375        6      1.229
```

© 1989, SYSTAT, Inc.

A few of you may have noticed that this analysis is the same as that for a repeated measures design with subjects as PLOTS, groups as A, and trials as B. Since this method becomes unwieldy for a large number of plots (subjects), **MGLH** offers a more compact method for repeated measures analysis which we take up later in this chapter.

6.21 Latin square designs

A Latin square design imposes a pattern on treatments in a factorial design to save experimental effort or reduce within cell error. As in the nested design, not all combinations of the square and other treatments are measured, so the model lacks certain interaction terms between squares and treatments. **MGLH** can analyze these designs easily if an extra variable denoting the square is included in the file. Here is a fixed effects example from Neter, Wasserman, and Kutner (1985). The SQUARE variable is represented in the cells of the design. For simplicity, the dependent variable, RESPONSE, has been left out.

	DAY1	DAY2	DAY3	DAY4	DAY5
WEEK1	D	C	A	B	E
WEEK2	C	B	E	A	D
WEEK3	A	D	B	E	C
WEEK4	E	A	C	D	B
WEEK5	B	E	D	C	A

You would set up the data as shown below (the LATIN file), coding the grouping variable SQUARE so that A=1, B=2, C=3, D=4, and E=5.

© 1989, SYSTAT, Inc.

DAY	WEEK	SQUARE	RESPONSE
1	1	4	18
1	2	3	17
1	3	1	14
1	4	2	21
1	5	5	17
2	1	3	13
2	2	2	34
2	3	5	21
2	4	1	16
2	5	4	15
3	1	1	7
3	2	4	29
3	3	2	32
3	4	5	27
3	5	3	13
4	1	5	17
4	2	1	13
4	3	3	24
4	4	4	31
4	5	2	25
5	1	2	21
5	2	5	26
5	3	4	26
5	4	3	31
5	5	1	7

Now we're ready to do the analysis.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: RESPONSE
- Click [Indep'dent] and select the independent variables: DAY, WEEK, SQUARE
- Click [Categories] and select the categorical variables, specifying the number of levels for each: DAY (5 levels), WEEK (5 levels), SQUARE (5 levels)
- Click OK

```
Dep Var:RESPONSE      N:   25   Multiple R:  .931   Squared Multiple R:  .867

                    Analysis of Variance

Source      Sum-of-Squares    DF  Mean-Square     F-Ratio        P

DAY             82.000         4       20.500       1.306      0.323
WEEK           477.200         4      119.300       7.599      0.003
SQUARE         664.400         4      166.100      10.580      0.001

Error          188.400        12       15.700
```

© 1989, SYSTAT, Inc.

6.22 Crossover and changeover designs

In crossover designs an experiment is divided into periods and the treatment of a subject changes from one period to the next. Changeover studies often use designs similar to a Latin square. A problem with these designs is that there may be a residual or carry-over effect of a treatment into the following period. This can be minimized by extending the interval between experimental periods. This is not always feasible. Fortunately there are methods to assess the magnitude of any carry-over effects that may be present.

Two-period crossover designs can be analyzed as repeated-measures designs (see below). More complicated crossover designs can also be analyzed by SYSTAT and carry-over effects can be assessed. Cochran and Cox (1957) present a study of milk production by cows under three different feed schedules, A (roughage), B (limited grain) and C (full grain). The design of the study has the form of two 3×3 Latin squares:

	COW					
	Latin square 1			Latin square 2		
Period	I	II	III	IV	V	V
1	A	B	C	A	B	C
2	B	C	A	C	A	B
3	C	A	B	B	C	A

© 1989, SYSTAT, Inc.

The data are set up in the WILLIAMS data file as follows.

COW	PERIOD1	PERIOD2	FEED	RESIDUAL	MILK
1	1	0	1	0	38
1	2	0	2	1	25
1	3	0	3	2	15
2	1	0	2	0	109
2	2	0	3	2	86
2	3	0	1	3	39
3	1	0	3	0	124
3	2	0	1	3	72
3	3	0	2	1	27
4	0	1	1	0	86
4	0	2	3	1	76
4	0	3	2	3	46
5	0	1	2	0	75
5	0	2	1	2	35
5	0	3	3	1	34
6	0	1	3	0	101
6	0	2	2	3	63
6	0	3	1	2	1

Notice that there are two variables for the effect of PERIOD. PERIOD is nested within each Latin square (the periods for cows in one square are unrelated to the periods in the other). The variable RESIDUAL indicates the treatment of the preceding period. For the first period for each cow there is no preceding period. Code these as zeroes as we do the non-nested PERIOD effects.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: MILK
- Click [Indep'dent] and select the independent variables: COW, FEED, PERIOD1, PERIOD2, RESIDUAL
- Click [Categories] and select the categorical variables, specifying the number of levels for each: COW (6 levels), FEED (3 levels), PERIOD1 (3 levels), PERIOD2 (3 levels), RESIDUAL (3 levels)
- Click OK

© 1989, SYSTAT, Inc.

```
Dep Var:   MILK     N:   18   Multiple R:  .995   Squared Multiple R:  .990

                        Analysis of Variance

  Source   Sum-of-Squares   DF  Mean-Square    F-Ratio       P

     COW         3835.950    5      767.190     15.402     0.010
 PERIOD1         6027.556    2     3013.778     60.502     0.001
 PERIOD2         5461.556    2     2730.778     54.821     0.001
    FEED         2854.550    2     1427.275     28.653     0.004
Residual          616.194    2      308.097      6.185     0.060

   Error          199.250    4       49.813
```

There is a significant effect of feed on milk production and an insignificant residual or carry-over effect in this instance. To obtain a combined periods within squares term, follow the previous analysis with:

- Select **MGLH/Test...** from the **Stats** menu
- Specify PERIOD1 & PERIOD2 for [Effects;] remember the ampersand (&)
- Click OK

6.23 Analysis of covariance

Winer (1971) uses the following data (in the file COVAR) in an analysis of covariance in which X is the covariate and TREAT is the treatment.

TREAT	X	Y
1	3	6
1	1	4
1	3	5
1	1	3
1	2	4
1	1	3
1	4	6
2	4	8
2	5	9
2	5	7
2	4	9
2	3	8
2	1	5
2	2	7
3	3	6
3	2	7
3	2	7
3	3	7
3	4	8
3	1	5
3	4	7

© 1989, SYSTAT, Inc.

Before analyzing the data with an analysis of covariance model, be sure there is no significant interaction between the covariate and the treatment. The assumption of no interaction is often called the "homogeneity of slopes" assumption, because it is tantamount to saying that the slope of the regression line of the dependent variable onto the covariate should be the same in all cells of the design.

Homogeneity of slopes (parallelism)

This is easy to test with a preliminary model. Estimate this model with the interaction between treatment (TREAT) and covariate (X) in the model.

- Select **MGLH/General linear model...** from the Stats menu
- Select the [Dependent] variable: Y
- Click [Indep'dent] and select the independent variables: TREAT, X
- Click [Interactions] and select TREAT*X
- Click [Categories] and select TREAT, specifying 3 levels
- Click OK

```
Dep Var:      Y     N:   21   Multiple R:  .921   Squared Multiple R:  .849

                        Analysis of Variance

  Source   Sum-of-Squares   DF  Mean-Square    F-Ratio       P

   TREAT          6.693    2        3.346      5.210     0.019
       X         15.672    1       15.672     24.399     0.000
 TREAT*
       X          0.667    2        0.334      0.519     0.605

   Error          9.635   15        0.642
```

The probability value for the treatment by covariate interaction is .605, so the assumption of homogeneity of slopes is plausible.

Covariance analysis

Now we need to fit the usual analysis of covariance model, in which we have only covariate and treatment as predictors.

- Select **MGLH/Regression...** from the **Stats** menu
- Select the [Dependent] variable: Y
- Select the [Indep'dent] variables: TREAT, X
- Click OK

© 1989, SYSTAT, Inc.

```
Dep Var:      Y     N:  21   Multiple R:  .916   Squared Multiple R:  .839

                        Analysis of Variance

Source    Sum-of-Squares   DF  Mean-Square    F-Ratio       P

TREAT         16.932        2       8.466      13.970     0.000
X             16.555        1      16.555      27.319     0.000

Error         10.302       17       0.606
```

Now we see that the treatment adjusted for the covariate is significant. In passing, you should notice that the coefficient for the covariate is significant. If it were not, the analysis of covariance could be wasting a degree of freedom without reducing mean square error enough to help us.

If you are wild and crazy and ready to estimate your analysis of covariance immediately without testing the homogeneity of slopes assumption, you can use the [Covariates] option of **Fully factorial (M)ANOVA...** You would analyze the COVAR data this way:

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variable: Y
- Click [Factors] and select the factor: TREAT; specify 3 levels
- Click [Covariate] and select the covariate: X
- Click OK

The [Covariate] option for **Fully factorial (M)ANOVA...** adds one or more covariates to the model. For incomplete factorials and similar designs, you must use **General linear model...** to do analysis of covariance.

Saving adjusted cell means

To save adjusted cell means, click [Save file]. After specifying a filename and clicking [Save], you are asked to choose what type of file to save; click [Save adjusted cell means].

The variable in the file is ESTIMATE. For multivariate designs, this variable has subscripts for different dependent variables, e.g. ESTIMATE(1), ESTIMATE(2). The ESTIMATE is computed for every case, so there are duplicates for cases in the same cell. You can use the **Data/By groups...** and **Stats...** or use DATA to print these estimates. To merge the estimates with the original data, use the USE command in DATA.

© 1989, SYSTAT, Inc.

MGLH computes the adjusted cell means the same way it computes estimates when saving residuals. Coefficients for variables not selected as [Factors] (i.e. covariates), however, are multiplied by the mean value for each cell. The grand mean (CONSTANT) is included in computing the estimates.

Multiple covariates

You can have two or more covariates in a model. Just add them to the [Covariates] selection list. Be sure to select any grouping factors in the [Factors] selection list, so you can specify their numbers of levels.

6.24 Hotelling's T^2

You can use **General linear model...** to calculate Hotelling's T^2 statistic.

One sample test

To get a one-sample test for the variables X and Y, for instance, select both X and Y as [Dependent] variables and select only the CONSTANT term for the [Indep'dent] list. Then, test a hypothesis with **Test...**

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: X, Y
- Click [Indep'dent] to obtain a CONSTANT term
- Click OK

- Select **MGLH/Test...** from the **Stats** menu
- Click OK

The *F*-test for CONSTANT is the statistic you want. It is the same as the Hotelling's T^2 for the hypothesis that the population means for X and Y are zero.

You can also test against the hypothesis that the means of X and Y have particular nonzero values, for example 10 and 15, by using [D matrix].

- Select **Test...** from the **Stats** menu
- Click [D matrix] and enter the matrix:
  10 15
- Click OK

© 1989, SYSTAT, Inc.

Two sample test

For a two-sample test, you must provide a categorical independent variable that represents the two groups

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: X, Y
- Click [Indep'dent] and select the independent variable: GROUP
- Click [Categories] and select GROUP; specify 2 levels
- Click OK

- Select **MGLH/Test...** from the **Stats** menu
- Specify GROUP for [Effects]
- Click OK

Hotelling's T^2 is the simplest form of multivariate analysis of variance, discussed fully in the next section.

6.25 Multivariate analysis of variance

The following data (in the file MANOVA) comprise a hypothetical experiment on rats assigned randomly to one of three drugs. Weight loss in grams was observed for the first and second weeks of the experiment. The data were analyzed in Morrison (1976) with a two-way multivariate analysis of variance, or a two-way MANOVA.

SEX	DRUG	WEEK(1)	WEEK(2)
1	1	5	6
1	1	5	4
1	1	9	9
1	1	7	6
1	2	7	6
1	2	7	7
1	2	9	12
1	2	6	8
1	3	21	15
1	3	14	11
1	3	17	12
1	3	12	10
2	1	7	10
2	1	6	6
2	1	9	7
2	1	8	10
2	2	10	13
2	2	8	7
2	2	7	6
2	2	6	9
2	3	16	12
2	3	14	9
2	3	14	8
2	3	10	5

© 1989, SYSTAT, Inc.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: WEEK(1), WEEK(2)
- Click [Indep'dent] and select the independent variables: SEX, DRUG
- Click [Interactions] and select the interaction SEX*DRUG
- Click [Categories] and select the categories, specifying the number of levels for each: SEX (2 levels), DRUG (3 levels)
- Click OK

Notice that the only difference between an ANOVA and MANOVA model is that a MANOVA model has more than one dependent variable. We will see later, in Example 6.27, that the repeated measures ANOVA model can be expressed as a MANOVA as well.

You can also use **Fully factorial (M)ANOVA...** to set up the MANOVA model for complete factorials. You could get the same results as above by doing the following:

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variables: WEEK(1), WEEK(2)
- Click [Factors] and select the factors, specifying the number of levels for each: SEX (2 levels), DRUG (3 levels)
- Click OK

Fully factorial (M)ANOVA... derives the degrees of freedom from your [Factors] selections and the numbers of levels you specify.

```
Number of cases processed:    24

Dependent Variable Means

                  WEEK(1)     WEEK(2)
                    9.750       8.667

Estimates of Effects  B = (X'X)^-1 X'Y

                  WEEK(1)     WEEK(2)
CONSTANT            9.750       8.667
   SEX    1         0.167       0.167
  DRUG    1        -2.750      -1.417
  DRUG    2        -2.250      -0.167
   SEX    1
  DRUG    1        -0.667      -1.167
   SEX    1
  DRUG    2        -0.417      -0.417
```

© 1989, SYSTAT, Inc.

```
Squared Multiple Correlations
                WEEK(1)     WEEK(2)
                  0.770       0.378
```

Notice that each column of the **B** matrix is now assigned to a separate dependent variable. It is as if we had done two runs of an ANOVA. The numbers in the matrix are the analysis of variance effects estimates.

Extended output

If you choose [Show extended results] from **Data/Formats...**, you also get standardized estimates of effects (like standardized regression coefficients) and total and residual sum of squares and cross product matrices.

Testing hypotheses

With more than one dependent variable, we do not get a single ANOVA table; instead we must test each hypothesis separately. The **A matrix...** operates the way we've seen before: it contrasts independent variables. Here are three hypotheses. We choose [Show complete results] for the second hypothesis to illustrate the detailed output.

- Select **MGLH/Test...** from the **Stats** menu
- Specify SEX for [Effects]
- Click OK

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Click OK

- Select **MGLH/Test...** from the **Stats** menu
- Specify DRUG for [Effects]
- Click OK

- Select **Formats...** from the **Data** menu
- Choose [Show standard results]
- Click OK

- Select **MGLH/Test...** from the **Stats** menu
- Specify SEX*DRUG for [Effects], using an asterisk to indicate interaction
- Click OK

© 1989, SYSTAT, Inc.

Here are the results, collected:

Sex

```
Test for effect called:
                SEX

Univariate F tests

   Variable      SS       DF       MS          F          P
   WEEK(1)       0.667     1       0.667       0.127      0.726
      Error     94.500    18       5.250
   WEEK(2)       0.667     1       0.667       0.105      0.749
      Error    114.000    18       6.333

Multivariate test statistics

        Wilks' Lambda =    0.993
          F-Statistic =    0.064    DF =   2,  17    Prob =    0.938

        Pillai Trace  =    0.007
          F-Statistic =    0.064    DF =   2,  17    Prob =    0.938

Hotelling-Lawley Trace =    0.008
          F-Statistic =    0.064    DF =   2,  17    Prob =    0.938
```

Drug

```
Test for effect called:
               DRUG

Null Hypothesis Contrast AB

                WEEK(1)     WEEK(2)

      1          -2.750      -1.417
      2          -2.250      -0.167

                                -1
Inverse Contrast A(X'X)  A'

                    1           2

      1          0.083
      2         -0.042       0.083

                                                        -1   -1
Hypothesis Sum of Product Matrix  H = B'A'(A(X'X)  A')  AB

                WEEK(1)     WEEK(2)

   WEEK(1)      301.000
   WEEK(2)       97.500      36.333

Error Sum of Product Matrix G = E'E

                WEEK(1)     WEEK(2)

   WEEK(1)       94.500
   WEEK(2)       76.500     114.000
```


© 1989, SYSTAT, Inc.

```
Univariate F tests

    Variable        SS       DF        MS            F            P

    WEEK(1)      301.000      2      150.500       28.667        0.000
       Error      94.500     18        5.250
    WEEK(2)       36.333      2       18.167        2.868        0.083
       Error     114.000     18        6.333

Multivariate Test Statistics

          Wilks' Lambda =      0.169
            F-Statistic =     12.199     DF =   4,  34     Prob =      0.000

           Pillai Trace =      0.880
            F-Statistic =      7.077     DF =   4,  36     Prob =      0.000
  Hotelling-Lawley Trace =      4.640
            F-Statistic =     18.558     DF =   4,  32     Prob =      0.000

                  Theta =      0.821  S =   2, M =  -.5, N =  7.5 Prob =      0.000

Test of Residual Roots

  Roots  1 through  2
    Chi-square statistic =     36.491     DF =      4     Prob =      0.000

  Roots  2 through  2
    Chi-square statistic =      1.262     DF =      1     Prob =      0.261

Canonical Correlations

                         1           2

                     0.906       0.244

Dependent Variable Canonical Coefficients
Standardized by Conditional (Within Groups) Standard Deviations

                         1           2
       WEEK(1)       1.437      -0.352
       WEEK(2)      -0.821       1.231

Canonical Loadings (Correlations Between Conditional
Dependent Variables and Dependent Canonical Factors)

                         1           2

       WEEK(1)       0.832       0.555
       WEEK(2)       0.238       0.971
```

Sex by Drug

```
Test for effect called:
                SEX
          BY
               DRUG

Univariate F tests

    Variable        SS       DF        MS            F            P

    WEEK(1)       14.333      2        7.167        1.365        0.281
       Error      94.500     18        5.250
    WEEK(2)       32.333      2       16.167        2.553        0.106
       Error     114.000     18        6.333
```


© 1989, SYSTAT, Inc.


```
Multivariate Test Statistics

          Wilks' Lambda =      0.774
            F-Statistic =      1.159     DF =   4,  34     Prob =      0.346

           Pillai Trace =      0.227
            F-Statistic =      1.152     DF =   4,  36     Prob =      0.348

  Hotelling-Lawley Trace =      0.290
            F-Statistic =      1.159     DF =   4,  32     Prob =      0.347

                  Theta =      0.221  S =   2, M =  -.5, N =  7.5 Prob =      0.298
```

Default output

A few things about the output. The sometimes long matrix formulas make explicit the hypothesis being tested. For MANOVA, hypotheses are tested with sums of squares and cross products matrices. Before printing the multivariate tests, however, **MGLH** prints the univariate tests. Each of these *F* statistics is constructed in the same way as the ANOVA model. The sums of squares for hypothesis and error are taken from the diagonals of the respective sum of product matrices. The univariate *F* test for the WEEK(1) DRUG effect, for example, is computed from 301.0/2 over 94.5/18, or hypothesis mean square divided by error mean square.

The next statistics printed are for the multivariate hypothesis. Wilks' Lambda (likelihood ratio criterion) varies between 0 and 1. Schatzoff (1966) has tables for its percentage points. The following *F* statistic is Rao's approximate (sometimes exact) *F* statistic corresponding to the likelihood ratio criterion (see Rao, 1973, p. 556). Pillai's trace and its *F* approximation are taken from Pillai (1960). The Hotelling-Lawley trace and its *F* approximation are documented in Morrison (1976). The last statistic is the largest root criterion for Roy's union-intersection test (see Morrison, 1976). Charts of the percentage points of this statistic, found in Morrison and other multivariate texts, are taken from Heck (1960). The probability value printed for Theta is not an approximation. It is what you find in the charts. These multivariate test statistics and their approximations are the same ones printed by SAS and other multivariate mainframe packages, except for Theta, which has an exact probability value only in SYSTAT.

In the first hypothesis, the multivariate statistics all have the same value for the *F* approximation. This is because the approximation is exact when there are only two groups (see Hotelling's T^2 in Morrison, 1976). In these cases, Theta is not printed because it has the same probability value as the *F* statistics.

© 1989, SYSTAT, Inc.

Extended output

Since we chose [Show extended results] for the second hypothesis, we get additional material.

Bartlett's residual root (eigenvalue) test

The chi-square statistics follow Bartlett (1947). The probability value for the first chi-square statistic should correspond to that for the approximate multivariate F statistic in large samples. In small samples, they might be discrepant, in which case you should generally trust the F statistic more. The subsequent chi-square statistics are recomputed leaving out the first and later roots until the last root is tested. These are sequential tests, and should be treated with caution, but they can be used to decide how many dimensions (roots, canonical correlations) are significant. The number of significant roots corresponds to the number of significant p values in this ordered list.

Canonical coefficients

Information on the canonical variates is printed next. Dimensions with insignificant chi-square statistics in the prior tests should be ignored in general. Corresponding to each canonical correlation is a canonical variate, whose coefficients have been standardized by the within groups standard deviations. If you want to standardize these variates by the total (sample) standard deviations, click [Total]. The default option, used here, is [Within groups]. The [Total] option is generally used for canonical correlation analysis or multivariate regression when groups are not present to introduce covariation among variates.

Canonical loadings

Finally, the [canonical loadings] are printed. These are correlations and thus provide information different from the canonical coefficients. In particular, you can identify "suppressor" variables in the multivariate system by looking for differences in sign between the coefficients and the loadings (which is the case with these data). See Bock (1975) and Wilkinson (1975b, 1977) for interpretation of these variates.

6.26 Multivariate analysis of covariance

Multivariate analysis of covariance is a trivial extension of the univariate model discussed in a previous section. You use **General linear model...** the same way; just select the [Dependent] variables and select one or more covariate variables in the [Indep'dent] list. Don't forget to click [Categories] and specify the number of levels for each categorical independent variable.

© 1989, SYSTAT, Inc.

If you have a one-way or complete factorial design, you can use **Fully factorial (M)ANOVA...** and click [Covariates] to select covariate terms.

6.27 Multivariate profile analysis

Morrison (1976), Bock (1975), and others describe how to analyze growth curves and other repeated measures designs by treating the trials factor as a multivariate profile. In this example, six rats were weighed at the end of each of 5 weeks. The weights were set up in the RATS file like this:

WEIGHT(1)	WEIGHT(2)	WEIGHT(3)	WEIGHT(4)	WEIGHT(5)
1	4	5	10	10
2	9	10	7	7
3	6	7	8	8
4	7	9	8	9
3	6	7	8	9
2	3	5	7	7

Our hypothesis is that these weights change over time. Here is how to test it.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: WEIGHT(1), WEIGHT(2), WEIGHT(3), WEIGHT(4), WEIGHT(5)
- Click [Indep'dent] to include a CONSTANT term
- Click OK

This models a mean weight for the rats at each week.

```
Number of cases processed:     6

Dependent Variable Means

                WEIGHT(1)   WEIGHT(2)   WEIGHT(3)   WEIGHT(4)   WEIGHT(5)
                    2.500       5.833       7.167       8.000       8.333

                                    -1
Regression Coefficients  B = (X'X)  X'Y

                WEIGHT(1)   WEIGHT(2)   WEIGHT(3)   WEIGHT(4)   WEIGHT(5)
  CONSTANT          2.500       5.833       7.167       8.000       8.333

Multiple Correlations

                WEIGHT(1)   WEIGHT(2)   WEIGHT(3)   WEIGHT(4)   WEIGHT(5)
                    0.000       0.000       0.000       0.000       0.000
```

© 1989, SYSTAT, Inc.

Notice that the multiple correlations are zero because we are regressing the weights on a constant. Because of this, the regression coefficients are the same as the means of each week.

To test our hypothesis, we do the following:

- Select **MGLH/Test...** from the **Stats** menu
- Click [Profile] and choose [Poly]
- Click OK

The [Profile] option constructs a **C** matrix for the contrast **BC'=0**, where **B** is the matrix of regression coefficients. Earlier, we used contrasts of the form **AB=0** to test hypotheses on independent (predictor) variables. Here, we are working on the dependent variables. In this example, there is a constant only on the right hand side of the equation, so there is no role for the **A** matrix to play.

The [Poly] option for [Profile] generates equally spaced orthogonal polynomials. The box next to [Poly] is for specifying your own spacing for unequally spaced orthogonal models. We'll use this box in the next example. Since there are 5 weights, the polynomials can be up to the fourth degree. This means our **C** matrix is 4 rows (degree) by 5 columns (week). SYSTAT produces this automatically. As a check, however, the program prints the **C** matrix generated. Here is the output from the test:

```
Hypothesis.

C Matrix
```

	1	2	3	4	5
1	-0.632	-0.316	0.000	0.316	0.632
2	0.535	-0.267	-0.535	-0.267	0.535
3	-0.316	0.632	0.000	-0.632	0.316
4	0.120	-0.478	0.717	-0.478	0.120

© 1989, SYSTAT, Inc.


```
Univariate F tests
```

Variable	SS	DF	MS	F	P
1	114.817	1	114.817	38.572	0.002
Error	14.883	5	2.977		
2	18.107	1	18.107	7.061	0.045
Error	12.821	5	2.564		
3	1.350	1	1.350	0.678	0.448
Error	9.950	5	1.990		
4	0.193	1	0.193	0.225	0.655
Error	4.279	5	0.856		

```
Multivariate Test Statistics

        Wilks' Lambda =      0.011
      F-Statistic     =     43.007   DF =  4,  2   Prob =  0.023

        Pillai Trace  =      0.989
      F-Statistic     =     43.007   DF =  4,  2   Prob =  0.023

Hotelling-Lawley Trace =     86.014
      F-Statistic     =     43.007   DF =  4,  2   Prob =  0.023
```

Notice the **C** matrix. The rows are polynomials. Each Univariate *F* Test covers one polynomial. Here, the linear and quadratic trends are significant (F=38.572, F=7.061). The cubic and quartic trends are not significant (F=.678, F=.225). Finally, the overall multivariate test shows that the weights taken as a whole show a significant trend.

Unevenly spaced polynomials

Sometimes the underlying metric of the profiles is not evenly spaced. Let's assume for a moment that the fifth weight was measured after the tenth week instead of the fifth. In that case, the polynomials have to be adjusted for the uneven spacing.

- Select **MGLH/Test...** from the **Stats** menu
- Click [Profile] and choose [Poly]
  - Specify spacing: 1, 2, 3, 4, 10
- Click OK

Notice how the **C** Matrix in the output differs from the previous one. The last point has been spread out further to the right.

```
Hypothesis.

C Matrix
```

	1	2	3	4	5
1	-0.424	-0.283	-0.141	0.000	0.849
2	0.625	0.077	-0.344	-0.638	0.281
3	-0.442	0.573	0.397	-0.563	0.036
4	0.184	-0.621	0.710	-0.276	0.003

© 1989, SYSTAT, Inc.

```
Univariate F tests

  Variable      SS      DF      MS          F          P

         1    67.213     1    67.213     23.959      0.004
     Error    14.027     5     2.805
         2    62.283     1    62.283    107.867      0.000
     Error     2.887     5     0.577
         3     4.574     1     4.574      1.014      0.360
     Error    22.550     5     4.510
         4     0.396     1     0.396      0.802      0.412
     Error     2.470     5     0.494

Multivariate Test Statistics

        Wilks' Lambda =     0.011
          F-Statistic =    43.007     DF =   4,   2     Prob =     0.023

        Pillai Trace =      0.989
          F-Statistic =    43.007     DF =   4,   2     Prob =     0.023

Hotelling-Lawley Trace =    86.014
          F-Statistic =    43.007     DF =   4,   2     Prob =     0.023
```

The univariate *F* tests are different, but the multivariate test is unchanged, since it measures variation across all components.

Difference contrasts

If you do not want to use polynomials, you can specify a C matrix that contrasts adjacent weeks.

- Select **MGLH/Test...** from the **Stats** menu
- Choose [Diff] for the [Profile] option
- Click OK

Notice the **C** matrix that this command generates in the output below:

```
Hypothesis.

C Matrix

                1          2          3          4          5

  1         1.000     -1.000      0.000      0.000      0.000
  2         0.000      1.000     -1.000      0.000      0.000
  3         0.000      0.000      1.000     -1.000      0.000
  4         0.000      0.000      0.000      1.000     -1.000

Univariate F tests

  Variable      SS      DF      MS          F          P

         1    66.667     1    66.667     17.241      0.009
     Error    19.333     5     3.867
         2    10.667     1    10.667     40.000      0.001
     Error     1.333     5     0.267
         3     4.167     1     4.167      0.566      0.486
     Error    36.833     5     7.367
         4     0.667     1     0.667      2.500      0.175
     Error     1.333     5     0.267
```


© 1989, SYSTAT, Inc.


```
Multivariate Test Statistics

        Wilks' Lambda =     0.011
          F-Statistic =    43.007     DF =   4,   2     Prob =     0.023

        Pillai Trace =      0.989
          F-Statistic =    43.007     DF =   4,   2     Prob =     0.023

Hotelling-Lawley Trace =    86.014
          F-Statistic =    43.007     DF =   4,   2     Prob =     0.023
```

In this case, each of the Univariate *F* tests covers the significance of the difference between the adjacent weeks indexed by the **C** matrix. For example, the F=17.241 shows that the first and second weeks differ significantly (like doing a dependent groups t-test). Unlike polynomials, these contrasts are not orthogonal.

Summing effects

The [Sum] choice for the [Profile] option of **Test...** allows you to sum across weeks.

- Select **MGLH/Test...** from the **Stats** menu
- Choose [Sum] for the [Profile] option
- Click OK

In our example, this hypothesis isn't exactly earth shattering. We are testing whether the overall weight (across weeks) significantly differs from zero. Naturally, the *F* value is fantastically significant. Notice the **C** matrix generated. It is simply a set of 1's that in the equation **BC'=0** sum all the coefficients in **B**. In a group-by-trials design, this **C** Matrix is useful for pooling trials and analyzing group effects.

```
Hypothesis.

C Matrix

        1          2          3          4          5

    1.000      1.000      1.000      1.000      1.000

Test of Hypothesis

    Source      SS      DF      MS          F          P

Hypothesis  6080.167     1  6080.167    295.632      0.000
     Error   102.833     5    20.567
```

© 1989, SYSTAT, Inc.

Special contrasts

To test any arbitrary contrast effects between dependent variables you can use [C matrix], which has the same form (without a column for the CONSTANT) as [A matrix]. The following commands test a linear trend across five trials:

- Select **MGLH/Test...** from the **Stats** menu
- Click [C matrix]and enter the matrix:
  −2 −1 0 1 2
- Click OK

6.28 Unweighting means

Sometimes you wish to unweight the cell means when you test hypotheses in an ANOVA. Suppose you have an experiment in which a few rats died before its completion. You do not want the hypotheses tested to depend upon the differences in cell sizes (which are presumably random). Here is an example from Morrison (1976). The data (MOTHERS) are hypothetical profiles on 3 scales of mothers in each of 4 socioeconomic classes.

CLASS	SCALE(1)	SCALE(2)	SCALE(3)
1	19	20	18
1	20	21	19
1	19	22	22
1	18	19	21
1	16	18	20
1	17	22	19
1	20	19	20
1	15	19	19
2	12	14	12
2	15	15	17
2	15	17	15
2	13	14	14
2	14	16	13
3	15	14	17
3	13	14	15
3	12	15	15
3	12	13	13
4	8	9	10
4	10	10	12
4	11	10	10
4	11	7	12

Morrison analyzes these data with the multivariate profile model for repeated measures. Since the hypothesis of parallel profiles across classes is not rejected, we can test whether the profiles are level. That is, do the scales differ when we pool the classes together?

© 1989, SYSTAT, Inc.

Pooling unequal classes can be done by weighting each according to sample size or averaging the means of the subclasses. First let's look at the model. Notice in the output below that the dependent variable means differ from the CONSTANT. The CONSTANT in this case is a mean of the cell means rather than the mean of all the cases.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: SCALE(1), SCALE(2), SCALE(3)
- Click [Indep'dent] and select the independent variable: CLASS
- Click [Categories] and select the categorical variable, specifying the number of levels: CLASS (4 levels)
- Click OK

```
Number of cases processed:    21

Dependent Variable Means

                  SCALE(1)    SCALE(2)    SCALE(3)
                    14.524      15.619      15.857

                                  -1
Estimates of Effects  B = (X'X)  X'Y

                  SCALE(1)    SCALE(2)    SCALE(3)
CONSTANT            13.700      14.550      14.988
   CLASS    1       4.300       5.450       4.763
   CLASS    2       0.100       0.650      -0.787
   CLASS    3      -0.700      -0.550       0.012

Squared Multiple Correlations

                  SCALE(1)    SCALE(2)    SCALE(3)
                     0.816       0.917       0.858
```

Here is how we test the hypothesis of equality of scale parameters without weighting the cell means:

- Select **MGLH/Test...** from the **Stats** menu
- Choose [Diff] for the [Profile] option
- Specify [Effects]: CONSTANT
- Click OK

© 1989, SYSTAT, Inc.

```
Test for effect called:
                CONSTANT

C Matrix

                     1          2          3

          1      1.000     -1.000      0.000
          2      0.000      1.000     -1.000

Univariate F tests

    Variable       SS       DF       MS             F            P

           1     14.012      1      14.012          4.652        0.046
       Error     51.200     17       3.012
           2      3.712      1       3.712          1.026        0.325
       Error     61.500     17       3.618

Multivariate Test Statistics

          Wilks' Lambda =      0.564
            F-Statistic =      6.191     DF =   2,  16     Prob =      0.010

          Pillai Trace =       0.436
            F-Statistic =      6.191     DF =   2,  16     Prob =      0.010

  Hotelling-Lawley Trace =      0.774
            F-Statistic =      6.191     DF =   2,  16     Prob =      0.010
```

If we believe, as Morrison does, that the differences in cell sizes do not reflect population subclass proportions, then we need to weight the cell means to get a grand mean, namely:

$$8\mu_1 + 5\mu_2 + 4\mu_3 + 4\mu_4$$

Expressed in terms of our analysis of variance parameterization, this is:

$$8(\mu+\alpha_1) + 5(\mu+\alpha_2) + 4(\mu+\alpha_3) + 4(\mu+\alpha_4)$$

Since the sum of effects is zero for a classification, and since we do not have an independent estimate of CLASS4, this expression is equivalent to the following:

$$8(\mu+\alpha_1) + 5(\mu+\alpha_2) + 4(\mu+\alpha_3) + 4(\mu-\alpha_1-\alpha_2-\alpha_3)$$

© 1989, SYSTAT, Inc.

which works out to

$$21\mu + 4\alpha_1 + \alpha_2$$

or

$$21\mu + 4\alpha_1 + \alpha_2 + 0\alpha_3$$

We use [A matrix] to test this hypothesis.

- Select **MGLH/Test...** from the **Stats** menu
- Click [A matrix] and enter the matrix:
  21 4 1 0
- Choose [Diff] for the [Profile] option
- Click OK

```
Hypothesis.

A Matrix

          1          2          3          4

     21.000      4.000      1.000      0.000

C Matrix

                     1          2          3

          1      1.000     -1.000      0.000
          2      0.000      1.000     -1.000

Univariate F tests

    Variable       SS       DF       MS             F            P

           1     25.190      1      25.190          8.364        0.010
       Error     51.200     17       3.012
           2      1.190      1       1.190          0.329        0.574
       Error     61.500     17       3.618

Multivariate Test Statistics

          Wilks' Lambda =      0.501
            F-Statistic =      7.959     DF =   2,  16     Prob =      0.004

          Pillai Trace =       0.499
            F-Statistic =      7.959     DF =   2,  16     Prob =      0.004

  Hotelling-Lawley Trace =      0.995
            F-Statistic =      7.959     DF =   2,  16     Prob =      0.004
```

© 1989, SYSTAT, Inc.

This is the multivariate *F* statistic that Morrison gets. Incidentally, for these data we are inclined to prefer the weighted means analysis (the first one, which is the default for **MGLH**) because these differences in cell frequencies probably reflect population base rates. They are not random, like the dead rats experiment.

6.29 One-way repeated measures

Let's reanalyze the RATS data as a simple repeated measures design. To use the [Repeat] option, you must specify the number of levels for each factor. Our dependent variables have only one level, so we specify 1 level.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: WEIGHT(1), WEIGHT(2), WEIGHT(3), WEIGHT(4), WEIGHT(5)
- Click [Indep'dent] to obtain a CONSTANT term
- Specify 5 for [Repeat]
- Click OK

As an alternative to **General linear model...**, you can use **Fully factorial (M)ANOVA...**. Since we have no factors, the procedure generates only the constant in the model:

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variables: WEIGHT(1), WEIGHT(2), WEIGHT(3), WEIGHT(4), WEIGHT(5)
- Specify 5 for [Repeat]
- Click OK

The [Repeat] option treats the dependent variables in the model as a set of repeated measures. It generates both the multivariate profile analysis we computed in the last example and the classical univariate repeated measures analysis found in most analysis of variance textbooks. For the multivariate analysis, **MGLH** generates equally spaced polynomials, partitions the hypotheses according to the proper effects, and prints each successively. Univariate sums of squares are derived from the orthogonal components used in the multivariate calculations.

© 1989, SYSTAT, Inc.

We collected the interesting parts of the output (which follows) in a table:

Source	*SS*	*df*	*MS*	*F*
Within Subjects				
Trials	134.467	4	33.617	16.033
Residual	41.933	20	2.097	

If you choose [Show extended results] from **Data/Formats...**, you get extended output. Here is the standard output:

```
Number of cases processed:     6

Dependent Variable Means

              WEIGHT(1)   WEIGHT(2)   WEIGHT(3)   WEIGHT(4)   WEIGHT(5)
                  2.500       5.833       7.167       8.000       8.333
```

```
 Univariate and Multivariate Repeated Measures Analysis


**************************
* Within Subjects Effects *
**************************

Test for effect called:
                CONSTANT

Single Degree-of-Freedom Polynomial Contrasts

      Degree          SS        DF          MS           F           P

           1     114.817         1     114.817      38.572       0.002
       Error      14.883         5       2.977
           2      18.107         1      18.107       7.061       0.045
       Error      12.821         5       2.564
           3       1.350         1       1.350       0.678       0.448
       Error       9.950         5       1.990
           4       0.193         1       0.193       0.225       0.655
       Error       4.279         5       0.856

Univariate Repeated Measures F-test

      Source          SS        DF          MS           F           P

  Hypothesis     134.467         4      33.617      16.033       0.000
       Error      41.933        20       2.097
```

© 1989, SYSTAT, Inc.

```
Multivariate Test Statistics

        Wilks' Lambda =     0.011
          F-Statistic =    43.007     DF =   4,   2     Prob =     0.023

        Pillai Trace =      0.989
          F-Statistic =    43.007     DF =   4,   2     Prob =     0.023

Hotelling-Lawley Trace =    86.014
          F-Statistic =    43.007     DF =   4,   2     Prob =     0.023
```

6.30 Repeated measures with grouping factors

The [Repeat] option enables you to handle grouping factors automatically. Here is an example from Winer (1971). There are two grouping factors (ANXIETY and TENSION) and one trials factor. The data are entered in the file REPEAT1 as follows:

ANXIETY	TENSION	TRIAL(1)	TRIAL(2)	TRIAL(3)	TRIAL(4)
1	1	18	14	12	6
1	1	19	12	8	4
1	1	14	10	6	2
1	2	16	12	10	4
1	2	12	8	6	2
1	2	18	10	5	1
2	1	16	10	8	4
2	1	18	8	4	1
2	1	16	12	6	2
2	2	19	16	10	8
2	2	16	14	10	9
2	2	16	12	8	8

One grouping factor

Let's first analyze these data with only one grouping factor (ANXIETY).

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: TRIAL(1), TRIAL(2), TRIAL(3), TRIAL(4)
- Click [Indep'dent] and select the independent variable: ANXIETY
- Click [Categories] and select the categorical variable, specifying the number of levels: ANXIETY (2 levels)
- Specify 4 for [Repeat]
- Click OK

If you wish, you can use **Fully factorial (M)ANOVA...** instead:

© 1989, SYSTAT, Inc.

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variables: TRIAL(1), TRIAL(2), TRIAL(3), TRIAL(4)
- Click [Factors] and select the factor, specifying the number of levels: ANXIETY (2 levels)
- Specify 4 for [Repeat]
- Click OK

We collected some of the univariate results in the table below. The full output from the procedure follows.

Source	*SS*	*df*	*MS*	*F*
Between Subjects				
A (anxiety)	10.083	1	10.083	.590
Subj w/groups	170.917	10	17.092	
Within Subjects				
B (trials)	991.500	3	330.500	128.627
AB	8.417	3	2.806	1.092
B×Subj w/groups	77.083	30	2.569	

The multivariate results correspond to the table:

Source	*Multiv. F*	*df*	*P*
B (trials)	64.854	3,8	.000
AB	2.451	3,8	.138

There is a lot of output.

```
Number of cases processed:  12

Dependent Variable Means

    TRIAL(1)    TRIAL(2)    TRIAL(3)    TRIAL(4)

      16.500      11.500       7.750       4.250
```

© 1989, SYSTAT, Inc.

```
Univariate and Multivariate Repeated Measures Analysis

**************************
* Between Subjects Effects *
**************************

Test for effect called:
                ANXIETY

Test of Hypothesis

      Source        SS       DF        MS            F          P

  Hypothesis     10.083        1     10.083        0.590      0.460
       Error    170.917       10     17.092
```

So far, we have the between subjects effect (ANXIETY), which is not significant (F=.59). The program continues to generate output for the within subjects effects. First, we see TRIALS, which corresponds to a test for CONSTANT because TRIALS*CONSTANT is the same as TRIALS.

```
**************************
* Within Subjects Effects *
**************************

Test for effect called:
                CONSTANT

Single Degree-of-Freedom Polynomial Contrasts

      Degree        SS       DF        MS            F          P

           1    984.150        1    984.150      190.051      0.000
       Error     51.783       10      5.178
           2      6.750        1      6.750        4.154      0.069
       Error     16.250       10      1.625
           3      0.600        1      0.600        0.663      0.434
       Error      9.050       10      0.905

Univariate Repeated Measures F-test

      Source        SS       DF        MS            F          P

  Hypothesis    991.500        3    330.500      128.627      0.000
       Error     77.083       30      2.569

Multivariate Test Statistics

          Wilks' Lambda =      0.039
            F-Statistic =     64.854     DF =   3,    8     Prob =     0.000

          Pillai Trace =       0.961
            F-Statistic =     64.854     DF =   3,    8     Prob =     0.000

      Hotelling-Lawley Trace =  24.320
            F-Statistic =     64.854     DF =   3,    8     Prob =     0.000
```


© 1989, SYSTAT, Inc.

Now we get the TRIALS*ANXIETY test. Remember, we are still within subjects, so the following test involves trials. Don't confuse this test with the between groups ANXIETY test above.

```
Test for effect called:
                ANXIETY

Single Degree-of-Freedom Polynomial Contrasts

      Degree        SS       DF        MS            F          P

           1     1.667        1      1.667        0.322      0.583
       Error    51.783       10      5.178
           2     3.000        1      3.000        1.846      0.204
       Error    16.250       10      1.625
           3     3.750        1      3.750        4.144      0.069
       Error     9.050       10      0.905

Univariate Repeated Measures F-test

      Source        SS       DF        MS            F          P

  Hypothesis     8.417        3      2.806        1.092      0.368
       Error    77.083       30      2.569

Multivariate Test Statistics

          Wilks' Lambda =      0.521
            F-Statistic =      2.451     DF =   3,    8     Prob =     0.13

          Pillai Trace =       0.479
            F-Statistic =      2.451     DF =   3,    8     Prob =     0.13

      Hotelling-Lawley Trace =  0.919
            F-Statistic =      2.451     DF =   3,    8     Prob =     0.13
```

Two grouping factors

Now let's analyze the design with two grouping factors (ANXIETY and TENSION). Remember to select grouping factors as [Indep'dent] variables and the trials factors as [Dependent] variables. The model includes an interaction between the grouping factors (ANXIETY*TENSION) also. Don't forget the [Repeat] option, or **MGLH** will assume you are doing a standard multivariate analysis of variance (MANOVA).

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: TRIAL(1), TRIAL(2), TRIAL(3), TRIAL(4)
- Click [Indep'dent] and select the independent variables: ANXIETY, TENSION
- Click [Interactions] and select the interaction: ANXIETY*TENSION
- Click [Categories] and select the categorical variables, specifying the number of levels for each variable: ANXIETY (2 levels) and TENSION (2 levels)
- Specify 4 for [Repeat]
- Click OK

© 1989, SYSTAT, Inc.

Because this is a complete factorial, you could use Fully factorial (M)ANOVA... instead. In this case, it is less work than General linear model...

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variables: TRIAL(1), TRIAL(2), TRIAL(3), TRIAL(4)
- Click [Factors] and select the factors, specifying the number of levels for each: ANXIETY (2 levels), TENSION (2 levels)
- Specify 4 for [Repeat]
- Click OK

We collected some of the univariate results in a table.

Source	*SS*	*df*	*MS*	*F*
Between Subjects				
A (anxiety)	10.083	1	10.083	.978
B (tension)	8.333	1	8.333	.808
AB	80.083	1	80.083	7.766
Subj w/groups	82.500	8	10.313	
Within Subjects				
C (trials)	991.500	3	330.500	152.051
AC	8.417	3	2.806	1.291
BC	2.167	3	4.056	1.866
ABC	12.750	3	4.250	1.955
C×Subj w/groups	52.167	24	2.174	

We also collected some of the multivariate results in a table:

Source	*Multiv. F*	*df*	*P*
C (trials)	127.686	3,6	.000
AC	6.183	3,6	.029
BC	3.546	3,6	.088
ABC	4.099	3,6	.067

The error terms for the multivariate tests are matrices, not individual values.

© 1989, SYSTAT, Inc.

The full output follows. First, we get between subjects effects in the order selected.

```
Number of cases processed:    12

Dependent Variable Means

                TRIAL(1)    TRIAL(2)    TRIAL(3)    TRIAL(4)
                  16.500      11.500       7.750       4.250

Univariate and Multivariate Repeated Measures Analysis

**************************
* Between Subjects Effects *
**************************

Test for effect called:
                ANXIETY

Test of Hypothesis

      Source        SS       DF        MS          F          P

  Hypothesis    10.083        1    10.083      0.978      0.352
       Error    82.500        8    10.313

Test for effect called:
                TENSION

Test of Hypothesis

      Source        SS       DF        MS          F          P

  Hypothesis     8.333        1     8.333      0.808      0.395
       Error    82.500        8    10.313

Test for effect called:
                ANXIETY
             BY
                TENSION

Test of Hypothesis

      Source        SS       DF        MS          F          P

  Hypothesis    80.083        1    80.083      7.766      0.024
       Error    82.500        8    10.313
```

© 1989, SYSTAT, Inc.

Next SYSTAT prints the within subjects effects. First, TRIALS.

```
*************************
* Within Subjects Effects *
*************************

Test for effect called:
                CONSTANT

Single Degree-of-Freedom Polynomial Contrasts

      Degree        SS        DF        MS             F             P

           1     984.150       1     984.150       247.845         0.000
       Error      31.767       8       3.971
           2       6.750       1       6.750         3.411         0.102
       Error      15.833       8       1.979
           3       0.600       1       0.600         1.051         0.335
       Error       4.567       8       0.571

Univariate Repeated Measures F-test

      Source        SS        DF        MS             F             P

  Hypothesis     991.500       3     330.500       152.051         0.000
       Error      52.167      24       2.174

Multivariate Test Statistics

          Wilks' Lambda =       0.015
            F-Statistic =     127.686     DF =  3,   6     Prob =     0.000

           Pillai Trace =       0.985
            F-Statistic =     127.686     DF =  3,   6     Prob =     0.000

  Hotelling-Lawley Trace =      63.843
            F-Statistic =     127.686     DF =  3,   6     Prob =     0.000
```

Now SYSTAT prints TRIALS*ANXIETY. Again, don't confuse this with the test for ANXIETY in the between groups section above. All remaining tests are within groups and involve an interaction with TRIALS.

```
Test for effect called:
                ANXIETY

Single Degree-of-Freedom Polynomial Contrasts

      Degree        SS        DF        MS             F             P

           1       1.667       1       1.667         0.420         0.535
       Error      31.767       8       3.971
           2       3.000       1       3.000         1.516         0.253
       Error      15.833       8       1.979
           3       3.750       1       3.750         6.569         0.033
       Error       4.567       8       0.571
```


© 1989, SYSTAT, Inc.


```
Univariate Repeated Measures F-test

      Source        SS        DF        MS             F             P

  Hypothesis       8.417       3       2.806         1.291         0.300
       Error      52.167      24       2.174

Multivariate Test Statistics

          Wilks' Lambda =       0.244
            F-Statistic =       6.183     DF =  3,   6     Prob =     0.029

           Pillai Trace =       0.756
            F-Statistic =       6.183     DF =  3,   6     Prob =     0.029

  Hotelling-Lawley Trace =       3.091
            F-Statistic =       6.183     DF =  3,   6     Prob =     0.029
```

Next we get TRIALS*TENSION.

```
Test for effect called:
                TENSION

Single Degree-of-Freedom Polynomial Contrasts

      Degree        SS        DF        MS             F             P

           1      10.417       1      10.417         2.623         0.144
       Error      31.767       8       3.971
           2       0.083       1       0.083         0.042         0.843
       Error      15.833       8       1.979
           3       1.667       1       1.667         2.920         0.126
       Error       4.567       8       0.571

Univariate Repeated Measures F-test

      Source        SS        DF        MS             F             P

  Hypothesis      12.167       3       4.056         1.866         0.162
       Error      52.167      24       2.174

Multivariate Test Statistics

          Wilks' Lambda =       0.361
            F-Statistic =       3.546     DF =  3,   6     Prob =     0.088

           Pillai Trace =       0.639
            F-Statistic =       3.546     DF =  3,   6     Prob =     0.088

  Hotelling-Lawley Trace =       1.773
            F-Statistic =       3.546     DF =  3,   6     Prob =     0.088
```

© 1989, SYSTAT, Inc.

Finally, SYSTAT calculates TRIALS*ANXIETY*TENSION.

```
Test for effect called:
                ANXIETY
             BY
                TENSION

Single Degree-of-Freedom Polynomial Contrasts

     Degree        SS       DF        MS             F            P

          1       9.600      1       9.600         2.418        0.159
      Error      31.767      8       3.971
          2       0.333      1       0.333         0.168        0.692
      Error      15.833      8       1.979
          3       2.817      1       2.817         4.934        0.057
      Error       4.567      8       0.571

Univariate Repeated Measures F-test

     Source        SS       DF        MS             F            P

 Hypothesis      12.750      3       4.250         1.955        0.148
      Error      52.167     24       2.174

Multivariate Test Statistics

          Wilks' Lambda =       0.328
            F-Statistic =       4.099     DF =   3,   6     Prob =       0.067

           Pillai Trace =       0.672
            F-Statistic =       4.099     DF =   3,   6     Prob =       0.067

  Hotelling-Lawley Trace =       2.050
            F-Statistic =       4.099     DF =   3,   6     Prob =       0.067
```

6.31 Repeated measures with two trial factors

The [Repeat] option enables you to handle two trials factors. Remember, you must specify the number of levels in each trial factor.

Here is an example from Winer (1971). The experiment has one grouping factor (NOISE) and two trials factors (PERIODS and DIALS). The trials factors must be sorted into a set of dependent variables (one for each pairing of the two factors' groups); it is useful to label the levels with a convenient mnemonic. We set the file up with variables P1D1 through P3D3. For example, P1D2 means a score in the (PERIODS=1, DIALS=2) cell. The data are in the file REPEAT2.

NOISE	P1D1	P1D2	P1D3	P2D1	P2D2	P2D3	P3D1	P3D2	P3D3
1	45	53	60	40	52	57	28	37	46
1	35	41	50	30	37	47	25	32	41
1	60	65	75	58	54	70	40	47	50
2	50	48	61	25	34	51	16	23	35
2	42	45	55	30	37	43	22	27	37
2	56	60	77	40	39	57	31	29	46

© 1989, SYSTAT, Inc.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: P1D1, P1D2, P1D3, P2D1, P2D2, P2D3, P3D1, P3D2, P3D3
- Click [Indep'dent] and select the independent variable: NOISE
- Click [Categories] and select the categorical variable, specifying the number of levels: NOISE (2 levels)
- Specify 3,3 for [Repeat]
- Click OK

You could also do the analysis with **Fully factorial (M)ANOVA...**:

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variables: P1D1, P1D2, P1D3, P2D1, P2D2, P2D3, P3D1, P3D2, P3D3
- Click [Factors] and select the factor, specifying the number of levels: NOISE (2 levels)
- Specify 3,3 for [Repeat]
- Click OK

Notice that the [Repeat] option specifies that the two trials factors have 3 and 3 levels respectively. **MGLH** assumes the subscript of the first factor will vary slowest in the ordering of the dependent variables. From that indexing, it generates the proper main effects and interactions.

Again, we collected some of the univariate results in a table.

Source	*SS*	*df*	*MS*	*F*
Between Subjects				
A (noise)	468.167	1	468.167	.752
Subj w/groups	2491.111	4	622.778	
Within Subjects				
B (periods)	3722.333	2	1861.167	63.389
AB	333.000	2	166.500	5.671
B×Subj w/groups	234.889	8	29.361	
C (dials)	2370.333	2	1185.167	89.823
AC	50.333	2	25.167	1.907
C×Subj w/groups	105.556	8	13.194	

© 1989, SYSTAT, Inc.

BC	10.667	4	2.667	.336
ABC	11.333	4	2.833	.357
BC*Subj w/groups	127.111	16	7.944	

We also collected some of the multivariate results.

Source	*Multiv. F*	*df*	*P*
B (periods)	28.145	2,3	.01
AB	8.111	2,3	06
C (dials)	91.456	2,3	.00
AC	1.155	2,3	.42
BC	331.445	4,1	.04
ABC	581.875	4,1	.03

Now let's look at the actual output. First, the between subjects effects (just one, NOISE):

```
Number of cases processed:     6

Dependent Variable Means

                  P1D1        P1D2        P1D3        P2D1        P2D2
                48.000      52.000      63.000      37.167      42.167
                P2D3        P3D1        P3D2        P3D3
                54.167      27.000      32.500      42.500
```

```
Univariate and Multivariate Repeated Measures Analysis

**************************
* Between Subjects Effects *
**************************

Test for effect called:
                  NOISE

Test of Hypothesis

     Source        SS       DF        MS          F          P

 Hypothesis    468.167       1    468.167      0.752      0.435
      Error   2491.111       4    622.778
```

Now, the within subjects effects. First, the program generates the effects for the first trials factor (PERIODS, according to the way we indexed it). The first one is the PERIODS main effect.

© 1989, SYSTAT, Inc.


```
**************************
* Within Subjects Effects *
**************************

    .....................
    + First Trials Factor +
    .....................

Test for effect called:
                  CONSTANT

Single Degree-of-Freedom Polynomial Contrasts

     Degree        SS       DF        MS          F          P

          1  3721.000        1  3721.000     73.441      0.001
      Error   202.667        4    50.667
          2     1.333        1     1.333      0.166      0.705
      Error    32.222        4     8.056

Univariate Repeated Measures F-test

     Source        SS       DF        MS          F          P

 Hypothesis  3722.333        2  1861.167     63.389      0.000
      Error   234.889        8    29.361

Multivariate Test Statistics

          Wilks' Lambda =      0.051
            F-Statistic =     28.145     DF =  2,  3     Prob =     0.011

           Pillai Trace =      0.949
            F-Statistic =     28.145     DF =  2,  3     Prob =     0.011

  Hotelling-Lawley Trace =     18.764
            F-Statistic =     28.145     DF =  2,  3     Prob =     0.011
```

Now, PERIODS*NOISE:

```
Test for effect called:
                  NOISE

Single Degree-of-Freedom Polynomial Contrasts

     Degree        SS       DF        MS          F          P

          1   225.000        1   225.000      4.441      0.103
      Error   202.667        4    50.667
          2   108.000        1   108.000     13.407      0.022
      Error    32.222        4     8.056

Univariate Repeated Measures F-test

     Source        SS       DF        MS          F          P

 Hypothesis   333.000        2   166.500      5.671      0.029
      Error   234.889        8    29.361
```

© 1989, SYSTAT, Inc.

Multivariate Test Statistics

Wilks' Lambda =	0.156					
F-Statistic =	8.111	DF =	2,	3	Prob =	0.062
Pillai Trace =	0.844					
F-Statistic =	8.111	DF =	2,	3	Prob =	0.062
Hotelling-Lawley Trace =	5.407					
F-Statistic =	8.111	DF =	2,	3	Prob =	0.062

Next the program generates the effects for the second trials factor (DIALS). The first one printed is the DIALS main effect:

```
-----------------------
+ Second Trials Factor +
-----------------------
```

Test for effect called:
CONSTANT

Single Degree-of-Freedom Polynomial Contrasts

Degree	SS	DF	MS	F	P
1	2256.250	1	2256.250	241.741	0.000
Error	37.333	4	9.333		
2	114.083	1	114.083	6.689	0.061
Error	68.222	4	17.056		

Univariate Repeated Measures F-test

Source	SS	DF	MS	F	P
Hypothesis	2370.333	2	1185.167	89.823	0.000
Error	105.556	8	13.194		

Multivariate Test Statistics

Wilks' Lambda =	0.016					
F-Statistic =	91.456	DF =	2,	3	Prob =	0.002
Pillai Trace =	0.984					
F-Statistic =	91.456	DF =	2,	3	Prob =	0.002
Hotelling-Lawley Trace =	60.971					
F-Statistic =	91.456	DF =	2,	3	Prob =	0.002

Now DIALS*NOISE:

Test for effect called:
NOISE

Single Degree-of-Freedom Polynomial Contrasts

Degree	SS	DF	MS	F	P
1	6.250	1	6.250	0.670	0.459
Error	37.333	4	9.333		
2	44.083	1	44.083	2.585	0.183
Error	68.222	4	17.056		

© 1989, SYSTAT, Inc.

Univariate Repeated Measures F-test

Source	SS	DF	MS	F	P
Hypothesis	50.333	2	25.167	1.907	0.210
Error	105.556	8	13.194		

Multivariate Test Statistics

Wilks' Lambda =	0.565					
F-Statistic =	1.155	DF =	2,	3	Prob =	0.425
Pillai Trace =	0.435					
F-Statistic =	1.155	DF =	2,	3	Prob =	0.425
Hotelling-Lawley Trace =	0.770					
F-Statistic =	1.155	DF =	2,	3	Prob =	0.425

Finally, MGLH generates the tests for the interaction between the two trials factors. First, we get PERIODS*DIALS.

```
-----------------------
+ Trials Interaction +
-----------------------
```

Test for effect called:
CONSTANT

Single Degree-of-Freedom Polynomial Contrasts

Degree	SS	DF	MS	F	P
1	0.375	1	0.375	0.045	0.842
Error	33.333	4	8.333		
2	3.125	1	3.125	0.815	0.418
Error	15.333	4	3.833		
3	6.125	1	6.125	0.750	0.435
Error	32.667	4	8.167		
4	1.042	1	1.042	0.091	0.778
Error	45.778	4	11.444		

Univariate Repeated Measures F-test

Source	SS	DF	MS	F	P
Hypothesis	10.667	4	2.667	0.336	0.850
Error	127.111	16	7.944		

Multivariate Test Statistics

Wilks' Lambda =	0.001					
F-Statistic =	331.445	DF =	4,	1	Prob =	0.041
Pillai Trace =	0.999					
F-Statistic =	331.445	DF =	4,	1	Prob =	0.041
Hotelling-Lawley Trace =	1325.780					
F-Statistic =	331.445	DF =	4,	1	Prob =	0.041

© 1989, SYSTAT, Inc.

And last, PERIODS*DIALS*NOISE:

```
Test for effect called:
                NOISE

Single Degree-of-Freedom Polynomial Contrasts

     Degree       SS      DF       MS          F          P

          1     1.042      1     1.042      0.125      0.742
      Error    33.333      4     8.333
          2     0.125      1     0.125      0.033      0.865
      Error    15.333      4     3.833
          3     3.125      1     3.125      0.383      0.570
      Error    32.667      4     8.167
          4     7.042      1     7.042      0.615      0.477
      Error    45.778      4    11.444

Univariate Repeated Measures F-test

     Source       SS      DF       MS          F          P

 Hypothesis    11.333      4     2.833      0.357      0.836
      Error   127.111     16     7.944

Multivariate Test Statistics

          Wilks' Lambda =      0.000
            F-Statistic =    581.875     DF =   4,   1     Prob =     0.031

          Pillai Trace =       1.000
            F-Statistic =    581.875     DF =   4,   1     Prob =     0.031

  Hotelling-Lawley Trace =    2327.500
            F-Statistic =    581.875     DF =   4,   1     Prob =     0.031
```

6.32 Repeated measures with three or more trials factors

There is no specific command to analyze repeated measures with more than two trials factors. It is possible, however, with a little thought and some knowledge of [C matrix] to analyze such a design.

Two levels on each of three trials

Let's suppose that we have a four-way design (A×B×C×D) in which A, B, and C are trials (within subject) factors with two levels each, and D is a between subjects effect. To analyze this design, do the following (without the [Repeat] option):

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variables: REP1, REP2, REP3, REP4, REP5, REP6, REP7, REP8
- Click [Factors] and select the factor, specifying the number of levels: D (2 levels)
- Click OK

© 1989, SYSTAT, Inc.

REP1 through REP8 are the eight within-subject repeated measures. These measures are ordered by conditions A, B and C with the last factor (C) rotating fastest like the last digit of an odometer in a car:

Factor	REP1	REP2	REP3	REP4	REP5	REP6	REP7	REP8
A	1	1	1	1	2	2	2	2
B	1	1	2	2	1	1	2	2
C	1	2	1	2	1	2	1	2

You must construct a specific C matrix with orthogonal contrasts for each within-subjects effect you wish to test. Since there are two levels for each within-subject factor in this design, each within-subject effect can be represented by a single contrast. In all of the output that would be produced by the commands below, the test for CONSTANT refers to the TRIALS effect for the within-subjects factor defined by the C matrix contrast. The test for D is actually the test for D interacting with that within-subjects effect.

Specify a linear trend in A with [C matrix]. Thus the four trials for level one of the A factor are contrasted against the four trials for level 2 of A. Then use the [All] option for [Effects] to test both the A and the A by D effects:

- Select **MGLH/Test...** from the **Stats** menu
- Click [C matrix] and enter the matrix:
  -1 -1 -1 -1 1 1 1 1
- Click [All] for the [Effects] option
- Click OK

Similarly, you could specify a linear trend across B to test the B and B by D effects by using this C matrix and the same procedures as above.

```
-1  -1   1   1  -1  -1   1   1
```

Next, the following C matrix performs the tests for C and C by D:

```
-1   1   1   1  -1   1  -1   1
```

© 1989, SYSTAT, Inc.

Form interaction effects among the trial factors by multiplying the corresponding contrast elements. For example, to test the A by B and A by B by D interactions, use this **C** matrix:

```
 1   1  -1  -1  -1  -1   1   1
```

To test the A by C and A by C by D interactions, use:

```
 1  -1   1  -1  -1   1  -1   1
```

To test the B by C and B by C by D interactions, use:

```
 1  -1  -1   1   1  -1  -1   1
```

And to test the A by B by C and A by B by C by D interactions, use:

```
-1   1   1  -1   1  -1  -1   1
```

Finally, test the effect of D (the between-subjects factor) with either of these procedures:

- Select **MGLH/Test...** from the **Stats** menu
- Click [C matrix] and enter the matrix:
  1 1 1 1 1 1 1 1
- Specify D for [Effects] option
- Click OK

or

- Select **MGLH/Test...** from the **Stats** menu
- Choose [Sum] for the [Profile] option
- Specify D for [Effects] option
- Click OK

The [Sum] option for [Profile] generates a **C** matrix of all ones.

Three levels on each of three trials

To analyze a similar design where A, B, and C each have three levels, use a similar procedure. There are 27 combinations of A, B and C. We again enter the repeated measures values in order of the treatments with C varying fastest.

We would select REP1 through REP27 as dependent variables and D as a factor with 2 levels, as before.

© 1989, SYSTAT, Inc.

It takes two contrasts to represent the effect of a trial factor with three levels. Therefore we use two polynomial contrasts for each effect.The **C** matrix to test for the A and A by D effects would be:

```
-1  -1  -1  -1  -1  -1  -1  -1  -1      linear trend in A
 0   0   0   0   0   0   0   0   0
 1   1   1   1   1   1   1   1   1

 1   1   1   1   1   1   1   1   1      quadratic trend in A
-2  -2  -2  -2  -2  -2  -2  -2  -2
 1   1   1   1   1   1   1   1   1
```

Note that commas are used to continue long contrasts across three lines.

The **C** matrix to test for the B and B by D effects would be:

```
-1  -1  -1   0   0   0   1   1   1      linear trend in B
-1  -1  -1   0   0   0   1   1   1
-1  -1  -1   0   0   0   1   1   1

 1   1   1  -2  -2  -2   1   1   1      quadratic trend in B
 1   1   1  -2  -2  -2   1   1   1
 1   1   1  -2  -2  -2   1   1   1
```

The **C** matrix to test for the C and C by D effects would be:

```
-1   0   1  -1   0   1  -1   0   1      linear trend in C
-1   0   1  -1   0   1  -1   0   1
-1   0   1  -1   0   1  -1   0   1

 1  -2   1   1  -2   1   1  -2   1      quadratic trend in C
 1  -2   1   1  -2   1   1  -2   1
 1  -2   1   1  -2   1   1  -2   1
```

Interaction terms require you to multiply all the contrasts for one factor by all the contrasts for a second. Thus an A by B interaction effect would require four contrasts altogether, each with 27 terms.

© 1989, SYSTAT, Inc.

The test for D again requires a C matrix of all ones, but with 27 measures it is easier to use the [Sum] choice for the [Profile] option of **Test...**

- Select **MGLH/Test...** from the **Stats** menu
- Choose [Sum] for the [Profile] option
- Specify D for [Effects] option
- Click OK

Remember that all the tests of significance of trials factors with more than two levels are multivariate and that separate univariate *F* statistics test each trend (row of the C Matrix) separately. Since our trends are orthogonal, each test for trend is independent.

6.33 Repeated measures analysis of covariance

Repeated measures analysis of covariance works similarly if the covariate does not vary within subjects (i.e., across trials). To do repeated measures ANCOVA where the covariate varies within subjects, you would have to set up your model like a split plot, with a different record for each measurement.

Here is an example from Winer (1971). This design has two trials (DAY), one covariate (AGE), and one grouping factor (SEX). The data are in the file WINER.

SEX	AGE	DAY(1)	DAY(2)
1	3	10	8
1	5	15	12
1	8	20	14
1	2	12	6
2	1	15	10
2	8	25	20
2	10	20	15
2	2	15	10

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: DAY(1), DAY(2)
- Click [Indep'dent] and select the independent variables: SEX, AGE
- Click [Categories] and select the categorical variable, specifying the number of levels: SEX (2 levels)
- Specify 2 for the [Repeat] option
- Click OK

© 1989, SYSTAT, Inc.

Or, use **Fully factorial (M)ANOVA...**:

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variables: DAY(1), DAY(2)
- Click [Factors] and select the factor, specifying the number of levels: SEX (2 levels)
- Click [Covariates] and select the covariate: AGE
- Specify 2 for the [Repeat] option
- Click OK

```
Number of cases processed:     8

Dependent Variable Means

                  DAY(1)      DAY(2)

                  16.500      11.875
```

```
 Univariate and Multivariate Repeated Measures Analysis

*****************************
* Between Subjects Effects *
*****************************

Test for effect called:
                SEX

Test of Hypothesis

     Source        SS       DF        MS          F          P

 Hypothesis     44.492        1     44.492      3.629      0.115
      Error     61.298        5     12.260
```

```
Test for effect called:
                AGE

Test of Hypothesis

     Source        SS       DF        MS          F          P

 Hypothesis    166.577        1    166.577     13.587      0.014
      Error     61.298        5     12.260
```

© 1989, SYSTAT, Inc.

```
**************************
* Within Subjects Effects *
**************************

Test for effect called:
                  CONSTANT

Test of Hypothesis

      Source        SS       DF        MS          F          P

  Hypothesis    22.366        1    22.366      6.899      0.008
       Error     6.248        5     1.250

-------------------------------------------------------------------

Test for effect called:
                       SEX

Test of Hypothesis

      Source        SS       DF        MS          F          P

  Hypothesis     0.494        1     0.494      0.395      0.557
       Error     6.248        5     1.250

-------------------------------------------------------------------

Test for effect called:
                       AGE

Test of Hypothesis

      Source        SS       DF        MS          F          P

  Hypothesis     0.127        1     0.127      0.102      0.763
       Error     6.248        5     1.250
```

The *F* statistics for the covariate and its interactions, namely AGE (13.587) and DAY*AGE (.102), are not ordinarily published. They are useful for understanding the adjustment made by the covariate, however. We did not test the homogeneity of slopes assumption in this analysis. To be safe, we should have used the **MGLH/General linear model...** dialog box to add the interaction of SEX*AGE and test that it and its interactions with the other variables are not significant.

© 1989, SYSTAT, Inc.

If you want to test the homogeneity of slopes assumption, run the following model first:

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: DAY(1), DAY(2)
- Click [Indep'dent] and select the independent variables: SEX, AGE
- Click [Interactions] and select the interaction: SEX*AGE
- Click [Categories] and select the categorical variable, specifying the number of levels: SEX (2 levels)
- Specify 1 for the [Repeat] option
- Click OK

6.34 Covariance alternatives to repeated measures

Analysis of covariance offers an alternative to repeated measures in a pre-post design. You can use the pre-test as a covariate in predicting the post-test. Here is how you would do a two group pre-post design:

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: POST
- Click [Indep'dent] and select the independent variables: GROUP, PRE
- Click [Categories] and select the categorical variable, specifying the number of levels: GROUP (2 levels)
- Click OK

When using this design, be sure to check the homogeneity of slopes assumption. That is, do the following and check that the interaction term, GROUP*PRE, is not significant.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variable: POST
- Click [Indep'dent] and select the independent variables: GROUP, PRE
- Click [Interactions] and select the interaction: GROUP*PRE
- Click [Categories] and select the categorical variable, specifying the number of levels: GROUP (2 levels)
- Click OK

6.35 Issues in repeated measures analysis

As you have seen, repeated measures designs can be generated in **MGLH** with a single procedure. You need not worry about weighting cases in unbalanced designs or selecting error terms. The program does this automatically. The multivariate sections of the output may be strange, however, to users who have not had a statistics course recently. There are a few things you should keep in mind.

© 1989, SYSTAT, Inc.

1) The sums of squares for the univariate *F* tests are pooled across subjects within groups and their interactions with trials. This means that the traditional analysis method has highly restrictive assumptions. You must assume that the variances within cells are homogeneous and that the covariances across all pairs of cells are equivalent (compound symmetry). There are some mathematical exceptions to this requirement, but they rarely occur in practice. Furthermore, the compound symmetry assumption rarely holds for real data.

2) For this reason, you should always look at the multivariate *F* statistic corresponding to the traditional univariate one on the output. If they lead to different conclusions, you are usually safer trusting the multivariate statistic because it does not require the compound symmetry assumption.

3) Compound symmetry is ***not*** required for the validity of the Single Degree-of-freedom Polynomial Contrasts. These polynomials partition sums of squares into orthogonal components. You should routinely examine the magnitude of these sums of squares relative to the Hypothesis sum of squares for the corresponding Univariate Repeated Measures *F*-test when your trials are ordered on a scale. Take a look at the Within Subjects Effects output for Example 6.30. The linear trend across trials (Degree 1) has a sum of squares of 984.15 and the total hypothesis sum of squares is only 991.5. This means that the linear trend accounts for over 99 percent of the variability across the four trials. The linear trend also happens to be highly significant.

4) In summary, think of the **MGLH** repeated measures output as an expanded traditional ANOVA table. The effects are printed in the same order as they appear in Winer (1971) and other texts, but they include the single degree of freedom and multivariate tests to protect you from false conclusions. If you are satisfied that both are in agreement, you can delete the additional lines in the output file and you will be left with the table you know and love.

5) You can test any hypothesis after you have estimated a repeated measures design and examined the output with **Hypothesis...**. For example, you can use the [Metric] choice for the [Profile] option of **Test...** to test single degree of freedom components in an unevenly spaced design. Or, you can use the [Diff] choice to do post-hoc tests on adjacent trials.

© 1989, SYSTAT, Inc.

This example tests adjacent trials within the ANXIETY effect in the group-by-trials design we analyzed.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: TRIAL(1), TRIAL(2), TRIAL(3), TRIAL(4)
- Click [Indep'dent] and select the independent variables: ANXIETY, TENSION
- Click [Interactions] and select the interaction: ANXIETY*TENSION
- Click [Categories] and select the categorical variables, specifying the number of levels for each variable: ANXIETY (2 levels), TENSION (2 levels)
- Specify 4 for [Repeat]
- Click OK

- Select **MGLH/Test...** from the **Stats** menu
- Specify ANXIETY for [Effects]
- Choose [Diff] for the [Profile] option
- Click OK

If you want still other contrasts on the trials factor, use [C matrix]

- Select **MGLH/Test...** from the **Stats** menu
- Specify ANXIETY for [Effects]
- Click [C matrix] and enter the matrix:
  −3 −1 1 3
- Click OK

This hypothesis is equivalent to a linear contrast on the TRIAL*ANXIETY interaction. You get the same *F* statistic as the program prints in the first Single Degree-of-Freedom Polynomial Contrast for the Within Subjects test of TRIALS*ANXIETY.

And, if you are really picky, you can use [A matrix] and [C matrix] together to do contrasts within contrasts. To use [A matrix], you must keep in mind how **MGLH** codes dummy variables (discussed at the beginning of the Analysis of Variance section).

© 1989, SYSTAT, Inc.

6) The [Repeat] option does not handle more than two repeated measures factors. If you need to analyze more, you must use [C matrix] and construct contrasts yourself. You can find tables of integer orthogonal polynomials in many ANOVA textbooks or you can use the [Profile] option on some dummy data to generate the numbers you need. To keep things organized, use mnemonic variable labels.

7) If you have missing data, try to impute values before running the analysis. The repeated measures model cannot accommodate missing data the way factorial ANOVA does, because the error terms must be computed from SUBJECT*TRIAL interactions.

8) If you choose [Show extended results] from **Data/Formats...**, SYSTAT shows sum of squares and cross products matrices and other multivariate information when you use the [Repeat] option. The output is considerably longer, so we omitted it in these examples.

© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

Multivariate models Contents

Introduction	283
Usage	284
General linear model...	284
Options	285
Using an SSCP, covariance, or correlation matrix as input	285
Previewing your model	286
Saving residuals and other results	287
Cross validation	287
Regression without the constant	287
Test...	287
Options	288
Selecting effects for hypothesis tests	289
Comparisons and contrasts	289
Choosing prior probabilities	289
Standardizing canonical coefficients	290
Factoring the hypothesis or error matrix	290
Saving results	291
Show complete results	291
Examples	292
6.36 Discriminant analysis	292
Saving canonical scores	296
Classification table	297
Canonical variable plot	298
Prior probabilities	298
Cross validation	299
Stepwise discriminant analysis	299
6.37 Within groups principal components analysis	299
6.38 Canonical correlation analysis	304

© 1989, SYSTAT, Inc.

Multivariate models Introduction

MGLH estimates the multivariate general linear model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

where **Y** is a vector or matrix of dependent variables, **X** is a vector or matrix of independent variables, **B** is a vector or matrix of regression coefficients, and **E** is a vector or matrix of random errors. See Searle (1971), Winer (1971), Neter and Wasserman (1974), or Cohen and Cohen (1983) for details.

In all the multivariate models, **Y** is a matrix of continuous measures. The **X** matrix can be either continuous or categorical dummy variables according to the type of model. For *discriminant analysis*, **X** is a matrix of dummy variables, as in ANOVA. For *principal components* analysis, **X** is a single column of 1's, namely, a constant. For *canonical correlation*, **X** is usually a matrix of continuous right-hand variables (and **Y** is the matrix of left-hand variables).

Once the parameters of a model have been estimated, they can be tested by any general linear hypothesis of the following form:

$$\mathbf{ABC}' = \mathbf{D}$$

where **A** is a matrix of linear weights on coefficients across the independent variables (the rows of **B**), **C** is a matrix of linear weights on the coefficients across dependent variables (the columns of **B**), **B** is the matrix of regression coefficients or effects, and **D** is a null hypothesis matrix (usually a null matrix). For the multivariate models described in this section, the **C** matrix is an identity matrix and the **D** matrix is null. The **A** matrix can have several different forms, but these are all submatrices of an identity matrix, and are easily formed with the [Effects] option.

© 1989, SYSTAT, Inc.

Multivariate models Usage

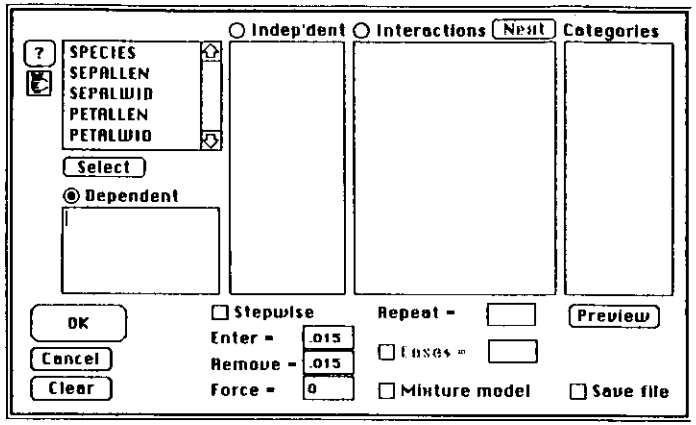
Multivariate models are generated through **MGLH/General linear model...** and tested with **Test...** The previous sections discuss these items in the context of regression and analysis of variance.

General linear model...

You can specify any multivariate linear model with **General linear model...** This dialog box has many options. You can ignore most of them for most multivariate models. Only three steps—marked with solid bullets below—are required. The examples below will show you which options are applicable to each model.

- Select MGLH/General linear model... from the Stats menu
- Select the [Dependent] variable(s)
  - Click [Indep'dent] and select the independent variables
  - Click [Interactions]and select the interaction terms, one at a time. for each interaction, select all the variables involved –SYSTAT links them with asterisks—and then click [Next] before selecting the variables involved in the next interaction term
  - Click [Categories] and select those independent variables that are categorical; for each, SYSTAT asks you to specify the number of levels: type the number and click OK. SYSTAT will then generate a set of dummy variables (using effects coding) for each of the categorical variables you have specified. There will be one fewer dummy variables than the number of levels you specify for a given categorical variable.
- Click OK

© 1989, SYSTAT, Inc.



Options

The following three options are more relevant to regression and analysis of variance:

[Stepwise]
Compute a stepwise regression, with [Enter] and [Remove] values, and [Force] criterion. See the "Regression" section.

[Repeat]
Create contrasts on dependent variables in multivariate regression. This technique is most frequently used in the context of repeated measures analysis of variance. See the "Analysis of variance" section.

[Mixture model]
Omit constant and treat predictors as mixtures that sum to a constant. See the "Regression" section.

Using an SSCP, covariance, or correlation matrix as input: [Cases]

Normally, you analyze raw cases by variables data with **MGLH.** You can, however, use a symmetric matrix data file (for example, a covariance matrix saved in a file from **Corr**) as input. If you use a matrix as input, you *must* use the [Cases] option to specify the sample size of the dataset that generated the matrix. The number you specify for [Cases] must be an integer greater than two.

© 1989, SYSTAT, Inc.

You can enter an SSCP, covariance, or correlation matrix by typing it into the Data Editor, by using DATA, or by saving it in a SYSTAT file and opening that file for **MGLH**. Be sure to include the dependent as well as independent variables in your matrix. **MGLH** picks out the dependent variable you name in your model.

SYSTAT uses the sample size to calculate degrees of freedom in hypothesis tests. **MGLH** determines the type of matrix (SSCP, covariance, etc.) and adjusts appropriately. With a correlation matrix, the raw and standardized coefficients are the same. Therefore, you cannot include a CONSTANT when using SSCP, covariance, or correlation matrices. Since these matrices are centered, the constant term has already been removed.

The triangular matrix input facility is useful for "meta-analysis" of published data and missing value computations. A few warnings, however. First, if you input correlation matrices from textbooks or articles, you may not get the same regression coefficients as those printed in the source. Because of round-off error, printed and raw data can lead to different results. Second, if you use pairwise deletion with **Corr**, the degrees of freedom for hypotheses will not be appropriate. You may not even be able to estimate the regression coefficients because of singularities.

In general, correlation matrices containing missing data produce coefficient estimates and hypothesis tests that are optimistic. You can correct for this specifying a sample size smaller than the number of actual observations (preferably set it equal to the smallest number of cases used for any pair of variables), but this is a crude guess you could refine only by doing Monte Carlo simulations. There is no simple solution. Beware, especially, of multivariate regressions (or MANOVA, etc.) with missing data on the dependent variables. You can usually compute coefficients, but hypothesis testing is particularly suspect.

Previewing your model

Click [Preview] to see your model displayed as an equation, with the dependent variables on the left and the independent variables (and any CONSTANT or interaction terms) on the right. Click OK to return to the **General linear model...** dialog box, where you can make any changes necessary. If you use the optional command interface, your MODEL statement is MODEL followed by the equation that you would see with [Preview].

© 1989, SYSTAT, Inc.

Saving residuals and other results

You can save residuals from any analysis by clicking [Save file]. A **Save as...** dialog box asks you to specify a filename, and after you click [Save], another dialog box lets you choose what gets saved and in what precision. See the "Regression" and "Analysis of variance" sections for further information on what gets saved. For multivariate analyses (discussed here), these residuals are rarely examined, because they are only marginal residuals for each separate dependent variable. See [Save file] below in **Test...** for information on other results saved for multivariate models, such as Mahalanobis distances, factor scores, canonical variable scores, and predicted group membership.

Cross validation

If you weight the data before computing a model, **MGLH** computes factor scores, Mahalanobis distances, canonical variable scores, and predicted group membership for cases with zero weight, even though they are not included in the cases used to estimate the model.

Regression without the constant

You can delete the constant from your model by selecting the CONSTANT term in the [Indep'dent] list and pressing the Delete key. You almost never want to delete the constant from multivariate models.

Test...

Test... lets you test parameters of the multivariate model estimated or factor the quadratic form of your model into orthogonal components. Linear hypotheses have the following form:

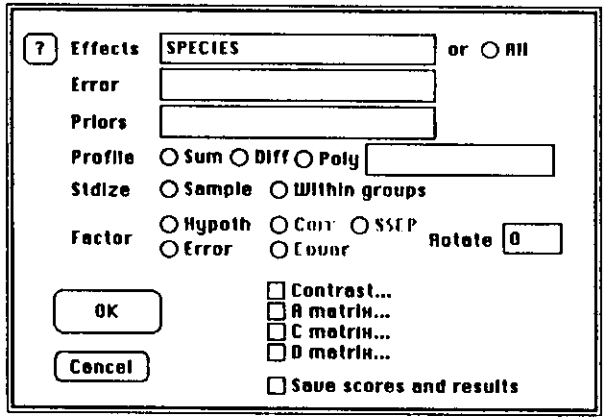
$$\mathbf{ABC'= D}$$

where **A** is a matrix of linear weights contrasting the coefficient estimates (the rows of **B**). **C** has as many columns as there are dependent variables. For most multivariate models, **C** is an identity matrix. Finally, **D** is a null hypothesis matrix (usually a null matrix). These matrices (**A**, **C**, and **D**) may be specified in several alternative ways and if they are not specified, they have default values.

© 1989, SYSTAT, Inc.

You can specify the **A**, **C**, and **D** matrices through the [Effects], [A matrix], [C matrix], [D matrix], [Contrast], and [Profile] options in the **Test...** dialog box. For most multivariate models, you will need only the [Effects] box.

- Select **MGLH/Test...** from the **Stats** menu
- Specify [Effects]; use asterisks to indicate interactions; use ampersands to test several effects simultaneously (see below)
- Click OK



Options

The following two options are more relevant to Regression and Analysis of Variance:

[Error]
Specify an error term. See the "Analysis of Variance" section.

[Profile]
Weight dependent variables with difference or polynomial contrasts. See the "Analysis of Variance" section.

© 1989, SYSTAT, Inc.

Selecting effects for hypothesis tests: [Effects]

The easiest way to specify one or more effects for testing in a hypothesis is to use the [Effects] option. For *discriminant analysis*, enter the name of your grouping variable in this box. For *principal components*, you don't need to use this box unless you have a grouping variable for within groups components. For *canonical correlation*, you need to enter all the right-hand variables in this box, with their names separated by ampersands (&).

Simply type in the name of the effect you want to test. If you have more than one effect in mind, then you must separate each effect name with an ampersand (&), e.g.

GROUP&REGION

If you wish to test an interaction term, then you must separate each variable name with an asterisk (*), e.g.

SEX*DRUG*GROUP

You may need to use both asterisks and ampersands. Suppose you wanted to test the effects A, B, and the interaction of A and B. You would type "A & B & A*B" for [Effects].

The [All] option is a shortcut if you wish to test all the effects in your model. This automatically generates separate tests for each effect in the model.

Comparisons and contrasts: [A matrix], [C matrix], [D matrix]

The [Contrast], [A matrix], [C matrix], and [D matrix] options are rarely used for multivariate models. Advanced users who need them should consult the "Regression" and "Analysis of variance" sections.

Choosing prior probabilities: [Priors]

[Priors] allows you to choose prior probabilities for *discriminant analysis* models. Enter your probabilities (one for each group) in the box separated by spaces. These probabilities should add to 1. For example, if you have three groups, priors might be .5, .3, and .2.

© 1989, SYSTAT, Inc.

Standardizing canonical coefficients: [Stdize]

Canonical coefficients in *discriminant analysis* are usually standardized by the pooled within groups standard deviations. This makes comparisons among variables easier when the measures are on different scales. For *canonical correlation* models, canonical coefficients are usually standardized by the total sample standard deviations. The default for [Stdize] is within sample standardizing, so be sure to click [Total] if you are doing canonical correlation.

Factoring the hypothesis or error matrix: [Factor]

Use the [Factor] options to do *principal components analysis* with **MGLH**. If you are just computing principal components on a single sample (no grouping variable), go directly to the **Factor** chapter and skip this section. While **MGLH** can compute ordinary components by using a model that contains only a constant, the **Factor** procedure gives you more detailed output. **MGLH** allows you, for example, to partial out effects due to grouping variables and to factor residual correlations. If between group variation is significant, the within group structure can differ substantially from the total structure (ignoring the grouping variable).

Choosing hypothesis or error matrix

If you have a model with grouping variables, choose [Hypoth] to factor the hypothesis (or between groups) matrix, and [Error] to factor the error (or within groups) matrix. Factoring the hypothesis matrix is an esoteric procedure, and is usually only of use when compared to the factoring of the error matrix. See Linn, Centra, and Tucker (1975) for an application of this technique.

Choosing type of matrix to factor

You can factor either sums of squares, covariances, or correlations. Choose the appropriate button next to [Factor].

Choosing number of components to rotate

Specify the number of components you wish to rotate with the [Rotate] option. Varimax rotation will be used. This option is ordinarily used for *principal components* analysis. For *canonical correlation*, [Rotate] can be used to rotate the dependent canonical factors. For *discriminant analysis*, [Rotate] can be used to rotate the canonical coefficients.

© 1989, SYSTAT, Inc.

Saving results

The [Save scores and results] option saves results from hypothesis tests. For *discriminant analysis*, this saves canonical variable scores (FACTOR), Mahalanobis distances to each group centroid (DISTANCE), posterior probabilities (PROB), original group membership (GROUP), and membership predicted by the model (PREDICT). For *principal components*, scores are saved (FACTOR). For *canonical correlation*, canonical scores for the left-hand set (FACTOR) are saved. If you want scores for the right-hand set, you need to refit the model by switching sets. See Example 6.38. If you click [Save scores and results], [Show extended results] is automatically chosen. This is because you generally want to see more detailed output when computing these statistics.

Show complete results

As with other items, **MGLH** produces extended output if you choose [Show extended results] from **Data/Formats...** The default is [Show standard results]. You get extended output from many of the routines until you change the setting in **Data/Formats...**

For estimation (with **MGLH/General Linear Model...**) , [Show extended results] adds the following output: total sum of product matrix, residual (or pooled within groups) sum of product matrix, residual (or pooled within groups) covariance matrix, and the residual (or pooled within groups) correlation matrix.

For hypothesis testing (with **MGLH/Test...**), [Show extended results] adds **A**, **C**, and **D** matrices, the matrix of contrasts, and the inverse of the cross products of contrasts, hypothesis and error sum of product matrices, tests of residual roots, canonical correlations, coefficients, and loadings. See the examples below for specific information on these statistics.

© 1989, SYSTAT, Inc.

Multivariate models Examples

6.36 Discriminant analysis

Here is an example using the IRIS data set from the Stats chapter. Fisher used this dataset to illustrate his discriminant function.

- Select **Formats...** from the **Stats** menu
- Choose [Show extended results]
- Click OK

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: SEPALLEN, SEPALWID, PETALLEN, PETALWID
- Click [Indep'dent] and select the independent variables: SPECIES
- Click [Categories] and select the categorical variable, specifying the number of levels: SPECIES (3 levels)
- Click OK

```
Number of cases processed:  150

Dependent Variable Means

                SEPALLEN    SEPALWID    PETALLEN    PETALWID
                   5.843       3.057       3.758       1.199

                                  -1
Estimates of Effects  B = (X'X)  X'Y

                SEPALLEN    SEPALWID    PETALLEN    PETALWID
CONSTANT           5.843       3.057       3.758       1.199
SPECIES    1      -0.837       0.371      -2.296      -0.953
SPECIES    2       0.093      -0.287       0.502       0.127

Standardized Estimates of Effects

                SEPALLEN    SEPALWID    PETALLEN    PETALWID
CONSTANT           0.000       0.000       0.000       0.000
SPECIES    1      -0.828       0.697      -1.066      -1.025
SPECIES    2       0.092      -0.540       0.233       0.136

Total Sum of Product Matrix

                SEPALLEN    SEPALWID    PETALLEN    PETALWID
SEPALLEN         102.168
SEPALWID          -6.323      28.307
PETALLEN         189.873     -49.119     464.325
PETALWID          76.924     -18.124     193.046      86.570
```


© 1989, SYSTAT, Inc.


```
Residual Sum of Product Matrix  E'E = Y'Y-Y'XB

                SEPALLEN    SEPALWID    PETALLEN    PETALWID
SEPALLEN          38.956
SEPALWID          13.630      16.962
PETALLEN          24.625       8.121      27.223
PETALWID           5.645       4.808       6.272       6.157

Residual Covariance Matrix  S
                             Y.X

                SEPALLEN    SEPALWID    PETALLEN    PETALWID
SEPALLEN          0.265
SEPALWID          0.093       0.115
PETALLEN          0.168       0.055       0.185
PETALWID          0.038       0.033       0.043       0.042

Residual Correlation Matrix  R
                              Y.X

                SEPALLEN    SEPALWID    PETALLEN    PETALWID
SEPALLEN          1.000
SEPALWID          0.530       1.000
PETALLEN          0.756       0.378       1.000
PETALWID          0.365       0.471       0.484       1.000

Squared Multiple Correlations

                SEPALLEN    SEPALWID    PETALLEN    PETALWID
                   0.619       0.401       0.941       0.929
```

The Residual Correlation and Residual Covariance matrices are the pooled within-groups matrices. We assume in computing a discriminant function that these are pooled across similar matrices within each group.

To compute the actual discriminant function, scores, and classification probabilities, we must test the null hypothesis that the groups are equivalent. Here is how to do it:

- Select **MGLH/Test...** from the **Stats** menu
- Specify SPECIES for [Effects]
- Click [Save scores and results]
- Click OK

© 1989, SYSTAT, Inc.

When you click the [Save scores and results] option, **MGLH** saves the canonical scores associated with the hypothesis. The scores are stored in subscripted variables named FACTOR. Because the [Effects] involve a categorical variable, the Mahalanobis distances (named DISTANCE) and posterior probabilities (named PROB) are saved into the same file. These distances are computed in the discriminant space itself. The closer a case is to a particular group's location in that space, the more likely it is that it belongs to that group. The probability of group membership is computed from these distances.

Here is the output:

```
Test for effect called:
                SPECIES

Null Hypothesis Contrast AB

                SEPALLEN    SEPALWID    PETALLEN    PETALWID

       1         -0.837       0.371      -2.296      -0.953
       2          0.093      -0.287       0.502       0.127

                    -1
Inverse Contrast A(X'X)  A'

                    1           2

       1          0.013
       2         -0.007       0.013

                                                  -1    -1
Hypothesis Sum of Product Matrix  H = B'A'(A(X'X)  A')  AB

                SEPALLEN    SEPALWID    PETALLEN    PETALWID

SEPALLEN          63.212
SEPALWID         -19.953      11.345
PETALLEN         165.248     -57.240     437.103
PETALWID          71.279     -22.933     186.774      80.413

Error Sum of Product Matrix G = E'E

                SEPALLEN    SEPALWID    PETALLEN    PETALWID

SEPALLEN          38.956
SEPALWID          13.630      16.962
PETALLEN          24.625       8.121      27.223
PETALWID           5.645       4.808       6.272       6.157
```


© 1989, SYSTAT, Inc.


```
Univariate F tests

    Variable          SS      DF          MS           F          P

SEPALLEN          63.212       2      31.606     119.265      0.000
   Error          38.956     147       0.265
SEPALWID          11.345       2       5.672      49.160      0.000
   Error          16.962     147       0.115
PETALLEN         437.103       2     218.551    1180.161      0.000
   Error          27.223     147       0.185
PETALWID          80.413       2      40.207     960.007      0.000
   Error           6.157     147       0.042

Multivariate Test Statistics

          Wilks' Lambda =      0.023
            F-Statistic =    199.145  DF =   8, 288     Prob =    0.000

          Pillai Trace =       1.192
            F-Statistic =     53.466  DF =   8, 290     Prob =    0.000

  Hotelling-Lawley Trace =     32.477
            F-Statistic =    580.532  DF =   8, 286     Prob =    0.000

                Theta =  0.970  S =  2, M =  .5, N =  71.0 Prob =    0.000

Test of Residual Roots

  Roots  1 through  2
    Chi-square statistic =    546.115  DF =     8     Prob =    .000

  Roots  2 through  2
    Chi-square statistic =     36.530  DF =     3     Prob =    .000

Canonical Correlations

                    1           2

                0.985       0.471

Dependent Variable Canonical Coefficients
Standardized by Conditional (Within Groups) Standard Deviations

                    1           2

SEPALLEN          0.427       0.012
SEPALWID          0.521       0.735
PETALLEN         -0.947      -0.401
PETALWID         -0.575       0.581

Canonical Loadings (Correlations Between Conditional
Dependent Variables and Dependent Canonical Factors)

                    1           2

SEPALLEN         -0.223       0.311
SEPALWID          0.119       0.864
PETALLEN         -0.706       0.168
PETALWID         -0.633       0.737
```

© 1989, SYSTAT, Inc.

Group Classification Function Coefficients

	1	2	3
SEPALLEN	23.544	15.698	12.446
SEPALWID	23.588	7.073	3.685
PETALLEN	-16.431	5.211	12.767
PETALWID	-17.398	6.434	21.079

Group Classification Constants

1	2	3
-86.308	-72.853	-104.368

Canonical scores have been saved

Our multivariate tests are all highly significant. If they were not, there would not be much point in proceeding further. The Dependent Variable Canonical Coefficients are used to produce discriminant scores. These coefficients are standardized by the within-groups standard deviations so you can compare their magnitude across variables with different scales. Since they are not raw coefficients, there is no need for a constant. The scores produced by these coefficients have overall zero mean and unit standard deviation within groups.

The group classification coefficients and constants comprise the Fisher discriminant functions for classifying the raw data. You can apply these coefficients to new data and assign each case to the group with the largest function value for that case.

Saving canonical scores

Let's look in the file of canonical scores produced with the [Save scores and results] option. First, we use the DATA command window to add a variable that labels cases so we can interpret plots. You can also do this with **Recode...** in the **Editor** menu.

```
DATA

USE "IRIS Work"
SAVE DISCRIM
LABEL GROUP/1 = 'A', 2 = 'B', 3 = 'C'
CASELIST
RUN
```


© 1989, SYSTAT, Inc.

Here are the first 5 cases of the output:

		FACTOR(1)	FACTOR(2)	DISTANCE(1)	DISTANCE(2)	DISTANCE(3)
		PROB(1)	PROB(2)	PROB(3)	GROUP	PREDICT
		GROUP$				
CASE	1	8.062	0.300	0.540	9.944	13.849
CASE	1	1.000	0.000	0.000	1.000	1.000
CASE	1	A				
CASE	2	7.129	-0.787	1.425	8.998	13.007
CASE	2	1.000	0.000	0.000	1.000	1.000
CASE	2	A				
CASE	3	7.490	-0.265	0.744	9.343	13.307
CASE	3	1.000	0.000	0.000	1.000	1.000
CASE	3	A				
CASE	4	6.813	-0.671	1.445	8.677	12.678
CASE	4	1.000	0.000	0.000	1.000	1.000
CASE	4	A				
CASE	5	8.132	0.514	0.772	10.046	13.923
CASE	5	1.000	0.000	0.000	1.000	1.000
CASE	5	A				

The variable GROUP is taken from the SPECIES variable we used in the analysis. PREDICT is the predicted value. Cases 1 through 5 have been classified correctly.

Classification table

Let's see how these assignments worked for the whole sample. We use **Tables** to tabulate actual group membership against predicted.

- Select **Tables/Tabulate...** from the **Stats** menu
- Select the variable GROUP from the selection list at the left
- Select the variable PREDICT from the right selection list
- Click OK

Table of GROUP (Rows) By PREDICT (Columns)

Frequencies

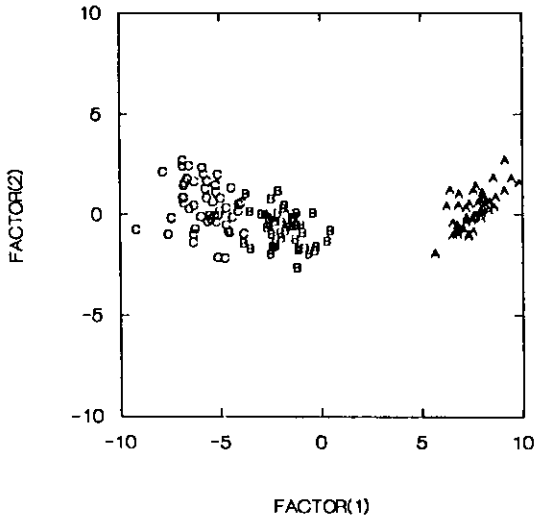
	1	2	3	Total
1	50	0	0	50
2	0	48	2	50
3	0	1	49	50
Total	50	49	51	150

© 1989, SYSTAT, Inc.

Canonical variable plot

The classification is almost perfect. Finally, we plot the scores in the discriminant space.

- Select **Plot/Plot** from the **Graph** menu
- Select the Y variable: FACTOR(2)
- Select the X variable: FACTOR(1)
- Click [Symb'l] and specify GROUP$ in the box next to the symbol choices
- Click OK



Notice that there is little variation on the second dimension. This is consistent with the output, which shows a large first canonical correlation and a much smaller second.

Prior probabilities

In this example, there were equal numbers of flowers in each group. Sometimes the probability of finding a case in each group is not the same across groups. You can adjust this prior probability by using the [Priors] option of **Test...**: specify the probabilities .5, .3, .2 in the box for [Priors].

© 1989, SYSTAT, Inc.

MGLH uses the probabilities you specify to compute the posterior probabilities that are saved into the file under the variable PROB. Be sure to specify a probability for each group in the category variable you choose for [Effects]. The probabilities should add to 1.

Cross validation

Sometimes you may want to classify new cases using a function computed on another ("training") set of cases. **Data/Weight...** enables this. Add a variable to your file with values of 1 for cases you want to compute the function on and 0 for the remaining cases. The predicted values are computed for all cases and you can check the success of the classification by using **Data/Select...** to select the 0 value of the weighting variable for **Tables**.

Some computer programs include these functions in a single "canned" procedure. That makes it easier to run simple examples. **MGLH** allows you to save discriminant scores and predicted values for any linear hypothesis, however. This way, you can include covariates and other types of moderators in your model.

Stepwise discriminant analysis

Finally, beware of stepwise discriminant analysis. All the cautions discussed in the stepwise regression section apply to the stepwise discriminant model, except they are more severe. While there are some Monte Carlo papers on the replication problem in stepwise regression, there is little information available on the stepwise discrimination problem. If you select two many discriminators out of a large set, it is unlikely to replicate in a new sample. Fishing expeditions are dangerous enough in the area of univariate statistics—in multivariate statistics, they are reckless. And remember, none of the statistics printed in stepwise programs (*F*-tests, Wilks' Lambda, etc.) have valid probabilities if you allow the computer to test more than one model. If you must remove variables from or add them to an equation, try another model in **MGLH** and look at the entire result carefully. Do not trust the computer to make your decisions for you. That's artificial intelligence.

6.37 Within groups principal components analysis

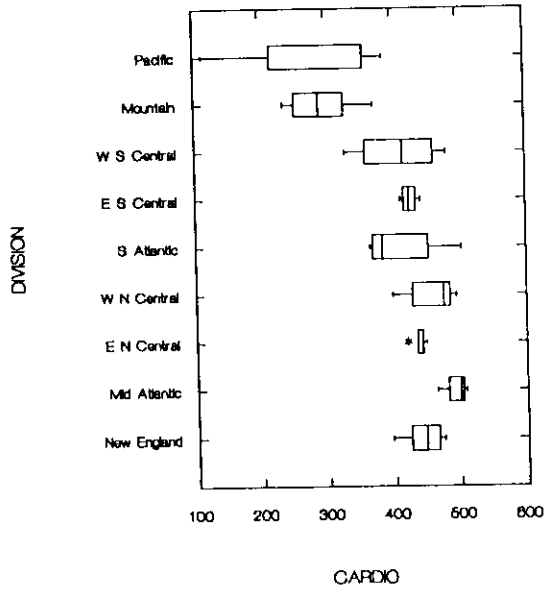
The following data (MEDICAL) are from the 1986 U.S. Statistical Abstract. They comprise death rates by cause from 9 census divisions of the country for that year. The divisions are in the next to the last column and the last column comprises the U.S. Post Office two letter state abbreviations. (These data are also in the USDATA data set.) We abbreviated the variable names PULMONAR, PNEU_FLU, and DIVISION to fit this table.

© 1989, SYSTAT, Inc.

ACCIDENT	CARDIO	CANCER	PULM	PNEU	DIABETES	LIVER	DIV	STATE$
37.7	466.2	213.8	33.6	21.1	15.6	14.5	1.0	ME
35.9	395.9	182.2	29.6	20.1	17.6	10.4	1.0	NH
41.3	433.1	188.1	33.1	24.0	15.6	13.1	1.0	VT
31.1	460.6	219.0	24.9	29.7	16.0	13.0	1.0	MA
28.6	474.1	231.5	27.4	17.7	26.2	13.4	1.0	RI
35.3	423.8	205.1	23.2	22.4	15.4	11.7	1.0	CT
31.5	499.5	209.9	23.9	26.0	17.1	17.7	2.0	NY
32.2	464.7	216.3	23.3	19.9	17.3	14.2	2.0	NJ
34.9	508.7	223.6	27.0	20.1	20.4	12.0	2.0	PA
33.2	443.1	198.8	27.4	18.0	18.9	10.2	3.0	OH
37.7	435.7	184.6	27.2	18.6	17.2	8.4	3.0	IN
32.9	449.6	193.2	22.9	21.3	15.3	12.5	3.0	IL
34.3	420.9	182.3	24.2	18.7	14.8	13.7	3.0	MI
33.8	444.3	189.4	22.5	21.2	15.7	8.7	3.0	WI
35.7	398.3	174.0	23.4	25.6	13.5	8.1	4.0	MN
38.6	490.1	199.1	31.2	28.3	16.6	7.9	4.0	IA
42.2	475.9	211.1	29.8	25.7	15.3	9.6	4.0	MO
48.2	401.0	173.7	18.2	25.9	14.9	7.4	4.0	ND
53.0	495.2	182.1	30.7	32.4	12.8	7.2	4.0	SD
40.8	479.6	187.4	31.6	28.3	13.5	7.8	4.0	NB
42.9	455.9	183.9	32.3	24.9	16.9	7.8	4.0	KS
38.8	404.5	202.8	25.3	16.0	25.0	10.5	5.0	DE
35.2	366.7	195.0	23.4	15.8	16.1	9.6	5.0	MD
37.4	365.3	174.4	22.4	20.3	11.4	9.2	5.0	VA
46.7	502.7	199.6	35.2	20.1	18.4	10.0	5.0	WV
45.4	392.6	169.2	22.6	19.8	13.1	10.2	5.0	NC
47.8	374.4	156.9	19.6	19.2	14.8	9.0	5.0	SC
48.2	371.4	157.9	22.6	20.5	13.2	10.4	5.0	GA
46.0	501.8	244.0	34.0	18.3	16.1	17.2	5.0	FL
48.8	442.5	194.7	29.8	22.9	15.9	9.1	6.0	KY
45.0	427.2	185.6	27.0	20.8	12.0	8.3	6.0	TN
48.9	411.5	185.8	25.5	16.8	16.1	9.1	6.0	AL
59.3	422.3	173.9	21.7	19.5	14.0	7.1	6.0	MS
51.0	482.0	202.1	29.0	22.7	15.0	8.7	7.0	AR
52.3	390.9	168.1	18.6	15.8	17.8	8.3	7.0	LA
62.5	441.4	182.4	27.6	24.5	15.3	9.6	7.0	OK
48.9	327.9	146.5	20.7	17.4	12.1	8.7	7.0	TX
59.0	372.2	170.7	33.4	25.1	14.4	11.1	8.0	MT
51.5	324.8	140.4	29.9	22.3	12.4	9.2	8.0	ID
67.6	264.2	112.2	27.7	18.5	9.2	9.2	8.0	WY
44.7	280.2	125.1	29.9	22.8	9.6	9.5	8.0	CO
62.3	235.6	137.2	28.7	17.8	17.5	13.1	8.0	NM
48.3	331.5	165.6	36.3	21.2	12.6	13.1	8.0	AZ
39.3	242.0	93.7	17.6	14.5	11.1	7.3	8.0	UT
57.3	299.5	162.3	32.3	13.7	11.1	15.4	8.0	NV
41.4	358.1	171.0	31.1	21.2	13.0	10.9	9.0	WA
41.6	387.8	179.4	33.8	23.1	11.2	10.4	9.0	OR
40.3	357.8	173.0	26.9	22.2	10.7	16.7	9.0	CA
85.8	114.6	76.1	8.3	12.4	3.4	11.0	9.0	AK
32.5	216.9	125.8	16.0	16.8	12.7	6.2	9.0	HI

© 1989, SYSTAT, Inc.

The variation in death rates between divisions in these data is substantial. Here is a grouped box plot of the second variable, CARDIO, by region. The variable DIVISION$ was created with DATA for labeling the plot. The other variables show similar regional differences.



If we analyze these data ignoring division, the correlations among death rates would be due substantially to between division differences. We might want to examine the pooled within-region correlations to see if the structure is different when divisional differences are statistically controlled. Accordingly, we will factor the residual correlation matrix after regressing medical variables onto an index variable denoting the census regions:

© 1989, SYSTAT, Inc.

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: ACCIDENT, CARDIO, CANCER, PULMONAR, PNEU_FLU, DIABETES, LIVER
- Click [Indep'dent] and select the independent variable: DIVISION
- Click [Categories] and select the categorical variable, specifying the number of levels: DIVISION (9 levels)
- Click OK

We omit the output from the estimation stage and proceed to the hypothesis stage. The following commands compute the principal components on the error (residual) correlation matrix and rotate the first two components to a varimax criterion.

- Select **MGLH/Test...** from the **Stats** menu
- Specify DIVISION for [Effects]
- Choose [Error] for the [Factor] option
- Click [Corr] to indicate the type of matrix being analyzed
- Specify 2 components to [Rotate]
- Click OK

Before looking at the output, let's examine the dialog box in more detail.

The [Factor] option may be used with any hypothesis. You can choose [Hypoth] or [Error] to choose which matrix—the hypothesis (**H**) matrix or the error (**G**) matrix—is to be factored. Ordinarily, when you test an hypothesis, the matrix product $\mathbf{G}^{-1}\mathbf{H}$ is factored and the latent roots of this matrix are used to construct the multivariate test statistic (see the references on MANOVA if this is unclear). The [Factor] option lets you compute principal components on **G** or **H** separately. Thus, it offers a direct way to compute principal components on residuals of any linear model you wish to fit. You can use any **A**, **C**, and/or **D** matrices in the hypothesis you are factoring, or any of the other commands that create these matrices.

We also clicked [Corr]. You can choose to factor the correlation matrix, the covariance matrix, or the error sum of squares and cross products matrix.

The [Rotate] option specifies how many components you want to rotate (by the varimax method). For other rotations, you have to use **Factor** in the **Stats** menu.

© 1989, SYSTAT, Inc.

Now we are ready to see the output. We've chosen [Show standard results] from **Data/Formats...** Notice the sorted, rotated loadings. Be careful. The indices on the side are not sorted. Only the loadings themselves are sorted. You can look back to the unsorted loadings to see which are which. The first component rotates to a dimension defined by all illness variables except diabetes, the second to one defined by accident versus diabetes.

Factoring Error Matrix

	1	2	3	4	5
1	1.000				
2	-0.286	1.000			
3	-0.345	0.852	1.000		
4	-0.180	0.838	0.758	1.000	
5	-0.120	0.575	0.335	0.434	1.000
6	-0.328	0.364	0.514	0.306	-0.144
7	0.112	0.352	0.522	0.383	0.092

	6	7
6	1.000	
7	0.030	1.000

Latent Roots

1	2	3	4	5
3.424	1.320	1.071	0.614	0.335

6	7
0.159	0.078

Loadings

	1	2	3	4	5
1	-0.378	0.627	0.409	0.539	0.063
2	0.944	0.086	-0.125	0.122	-0.058
3	0.939	-0.080	0.172	-0.049	0.023
4	0.879	0.140	0.005	0.172	-0.360
5	0.520	0.466	-0.630	0.054	0.320
6	0.478	-0.708	0.279	0.325	0.269
7	0.486	0.418	0.619	-0.411	0.154

	6	7
1	-0.055	0.017
2	-0.164	-0.200
3	-0.223	0.174
4	0.216	0.048
5	0.090	0.043
6	0.119	-0.020
7	0.101	-0.048

© 1989, SYSTAT, Inc.

```
Rotated Loadings on First   2 Principal Components

                    1          2

          1      -0.018      0.732
          2       0.863     -0.394
          3       0.775     -0.535
          4       0.833     -0.314
          5       0.683      0.147
          6       0.064     -0.852
          7       0.630      0.122

Sorted Rotated Loadings on First   2 Principal Components
(loadings less than .25 made 0.)

                    1          2

          1       0.863     -0.394
          2       0.833     -0.314
          3       0.775     -0.535
          4       0.683      0.000
          5       0.630      0.000
          6       0.000     -0.852
          7       0.000      0.732
```

You can compare these loadings with those for a factor analysis of the same data ignoring national regions in the **Factor** chapter. In that analysis diabetes loaded with the other illness variables and the second component is defined by PNEU_FLU. Thus systematic regional illness differences are sufficient to influence the loading pattern.

6.38 Canonical correlation analysis

Suppose you have 10 dependent variables MMPI(1)–MMPI(10) and 3 independent variables RATER(1)–RATER(3). Here's how you would get the canonical correlations and the dependent canonical coefficients.

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Click OK

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: MMPI(1), MMPI(2), ..., MMPI(10
- Click [Indep'dent] and select the independent variables: RATER(1), RATER(2), RATER(3)
- Click OK

- Select **MGLH/Test...** from the **Stats** menu
- Specify RATER(1) & RATER(2) & RATER(3) for [Effects], using ampersands
- Choose [Total] for the [Stdize] option
- Click OK

© 1989, SYSTAT, Inc.

The canonical correlations are displayed and, if you want, you can rotate the dependent canonical coefficients with the [Rotate] option.

If you want the coefficients for the independent variables, run **MGLH** again with the model reversed:

- Select **MGLH/General linear model...** from the **Stats** menu
- Select the [Dependent] variables: RATER(1), RATER(2), RATER(3)
- Click [Indep'dent] and select the independent variables: MMPI(1), MMPI(2), ..., MMPI(10)
- Click OK

And test the hypothesis the other way around:

- Select **MGLH/Test...** from the **Stats** menu
- Specify MMPI(1) & MMPI(2) & ... & MMPI(10) for [Effects], using ampersands
- Choose [Sample] for the [Stdize] option
- Click OK

© 1989, SYSTAT, Inc.

7 Nonlin

Introduction 308
Modeling the dose-response function 308
Loss functions 312
Maximum likelihood 312
Least absolute values 312
Usage 313
Model... 313
Estimation method 316
Initial values 316
Iterations 316
Convergence 316
Maximum likelihood estimation 317
Printing standard errors 317
Saving files 317
Loss function... 317
Resume 319
Show complete results 319
Operation 319
Asymptotic standard errors and parameter intercorrelations 319
Parameter bounds and model constraints 320
Problems 321
Computation 323
Missing data 323
Algorithms 323
Examples 325
7.1 Estimating a mean 325
7.2 Estimating a median 326
7.3 Linear regression 327
7.4 Least absolute values regression 328
7.5 Nonlinear regression 328
7.6 Piecewise regression 333
7.7 Maximum likelihood estimation 337
7.8 Logistic models—iteratively reweighted least squares 339
7.9 Logistic models—maximum likelihood estimation 340
7.10 Probit models—maximum likelihood estimation 342
7.11 Kinetic models 345
7.12 Deterministic models—minimizing an analytic function 348
7.13 A random model 349
7.14 Contouring the loss function 349

© 1989, SYSTAT, Inc.

Nonlin 7

Overview

Nonlin estimates parameters for a variety of nonlinear models. You use **Model...** to specify a general algebraic equation model. **Model...** uses the same expression parser as SYSTAT BASIC. You can use any function or operator available in SYSTAT BASIC except the multi-variable functions (e.g., SUM, AVG, etc.). In addition, **Loss function...** permits loss functions other than least squares. Maximum likelihood estimation, for example, can be accomplished by specifying negative log-likelihood with **Loss function...**

Nonlin uses two minimization methods: Quasi-Newton or Simplex. These allow efficient fitting of the greatest range of functions. [Save file] lets you saves residuals and estimates to analyze with other SYSTAT procedures.

© 1989, SYSTAT, Inc.

The following data are from a toxicity study for a drug designed to combat tumors. The table shows the proportion of laboratory rats dying (Response) at each dose level (Dose) of the drug. Clinical studies usually scale dose in natural logarithm units, which are listed in the center column (Log Dose). We arbitrarily set the Log Dose to –4 for zero Dose for the purpose of plotting and fitting with a linear model.

Dose	Log Dose	Response
0.00	-4.000	0.026
0.10	-2.303	0.120
0.25	-1.386	0.088
0.50	-0.693	0.169
1.00	0.000	0.281
2.50	0.916	0.443
5.00	1.609	0.632
10.00	2.303	0.718
25.00	3.219	0.820
50.00	3.912	0.852
100.00	4.605	0.879

Modeling the dose-response function

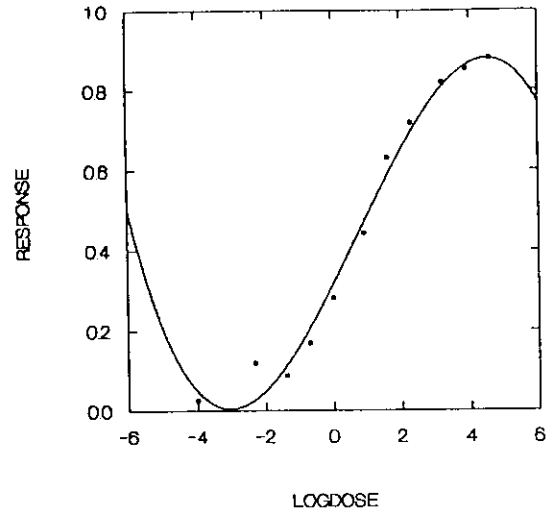
The plot of Response against Log Dose is clearly curvilinear. The S-shaped function suggests that we could use a linear model with linear, quadratic, and cubic terms (i.e. a polynomial function) to fit a curved line to the data. Here are the results from **MGLH**:

```
Dep Var:RESPONSE      N:   11    Multiple R:  .993    Squared Multiple R:  .986
Adjusted squared multiple R:  .981     Standard Error of Estimate:          0.047

  Variable     Coefficient     Std Error      Std Coef Tolerance     T    P(2 tail)

CONSTANT             0.314         0.021         0.000  .          15.317    0.000
 LOGDOSE             0.166         0.013         1.344  .1678077   12.486    0.000
 LOGDOSE*
 LOGDOSE             0.009         0.002         0.202  .7706010    4.027    0.005
 LOGDOSE*
 LOGDOSE*
 LOGDOSE            -0.004         0.001        -0.491  .1517832   -4.344    0.003
```

Notice that all the coefficients are highly significant and the overall fit is excellent (R^2=.986). Even the tolerances are relatively large, so we needn't worry about collinearity. The residual plots for this function are reasonably well behaved. There is no significant autocorrelation in the residuals.

© 1989, SYSTAT, Inc.

The following figure shows the observed data and the fitted curve.



How do the researchers interpret this plot? First of all, the curve is consistent with the printed output; it fits extremely well in the range of the data. Putting the fitted curve into ordinary language, we can say that fewer animals die at lower dosages and more at higher. At the extremes, however, more animals die with extremely low dosages and fewer animals die at extremely high dosages.

This is nonsense. While it is possible to imagine some drugs (arsenic, for example) for which dose-response functions are nonmonotonic, the model we fit makes no sense for a clinical drug of this sort. Second, the cubic function we fit extrapolates beyond the 0–1 response interval. It implies that there is something beyond dying and something less that living. Third, the parameters of the model we fit have no theoretical interpretation.

© 1989, SYSTAT, Inc.

Clinical researchers usually prefer to fit quantal response data like these with a bounded monotonic response function of the following form:

$$\text{proportion dying} = \alpha + \frac{1-\alpha}{1+e^{[\beta - \gamma \log(\text{dose})]}}$$

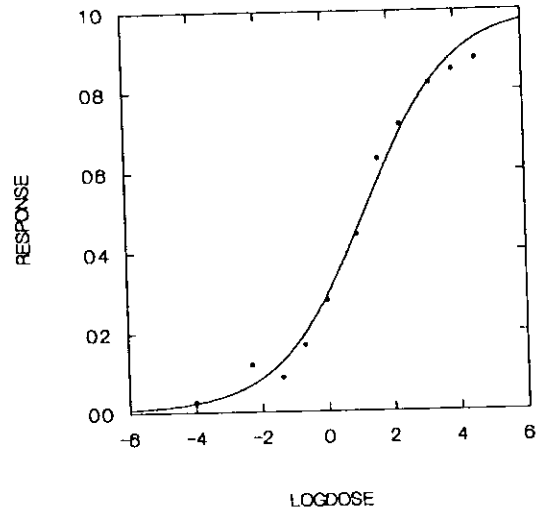
- α the background response, or rate of dying
- β a location parameter for the curve
- γ a slope parameter for the curve

Estimating a quantity called LD50 is the usual purpose of this type of study. LD50 is the Dose at which 50 percent of the animals are expected to die. LD50 is $e^{\beta/\gamma}(1-2\alpha)^{1/\gamma}$.

Notice how the parameters of this model make theoretical sense. We have a problem, however. We cannot fit an intrinsically nonlinear model like this with a linear regression program. We cannot even transform this equation, using logs or other mathematical operators, to a linear form. The cubic linear model we fit before was nonlinear in the data but linear in the parameters. As we saw in the **MGLH** introduction, linear models involve additive combinations of parameters. The model we want to fit now is nonlinear in the data *and* nonlinear in the parameters.

© 1989, SYSTAT, Inc.

We need a program that fits this type of model iteratively. **Nonlin** begins with initial estimates of parameter values and modifies them in small steps until the fit of the curve to the data is as close as possible. Here is the result:



Notice how the curve tapers at the ends so that it is bounded by 0 and 1 on the Response scale. This behavior fits our theoretical ideas about the effect of this drug. The value for LD50 is 3.295, which is in raw Dose units.

Interestingly, this model does not fit significantly better than the cubic polynomial. Both have comparable sums of squared residuals. True, the cubic model has four parameters and we have used only three. Nevertheless, this example should convince you that blind searching for models that produce good fits is not good science. It is even possible that a model with a poorer fit can be the true model generating data and one with a better fit can be bogus.

© 1989, SYSTAT, Inc.

Loss functions

Nonlinear estimation includes a broad variety of statistical procedures. We have performed nonlinear least squares, which is analogous to ordinary least squares (such as calculated in MGLH). Both methods minimize squared deviations of the dependent variable data values from values estimated by the function at the same independent variable data points. In these cases, LOSS is the sum of least squares.

Other types of LOSS functions can be defined which produce different estimates of parameters in the same functions. The most widely used LOSS is negative *log likelihood*. This LOSS is used for *maximum likelihood* estimation. Other loss functions are used for robust estimators and nonparametric procedures.

Maximum likelihood

A maximum likelihood estimate of a parameter is a value of that parameter in a given distribution that has the highest probability of generating the observed sample data. Sometimes maximum likelihood and least squares estimators coincide (as in fixed effects, fully crossed, balanced factorial ANOVA), and at other times they diverge. In our quantal response data example, the maximum likelihood estimates are different. They can be computed in **Nonlin** by using the **Loss...** function. See Example 7.7 for a general introduction and Example 7.8 for further information.

Least absolute values

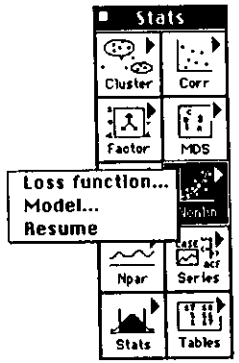
As an example of other types of loss functions, consider minimizing least absolute values of deviations of the dependent variable data values from values estimated by the function at the same independent variable data points. This procedure produces estimates which, on average, are influenced less by outliers than the least squares estimates. This is because squaring a large value increases its impact. While there are more sophisticated robust procedures, least absolute values estimates are easy to compute in **Nonlin** and fun to compare with least squares estimates. You can see how this is done in Example 7.4.

© 1989, SYSTAT, Inc.

Usage

Nonlin has three items. **Model...** lets you set up and estimate models. **Loss function...** lets you set up models other than least-squares. **Resume** resumes iterations where the **Model...** computations leave off.

Generally, you will only need to use **Model...** Use **Loss function...** to specify loss functions other than least squares. Use **Resume** after estimating to get more iterations if your model has failed to converge.



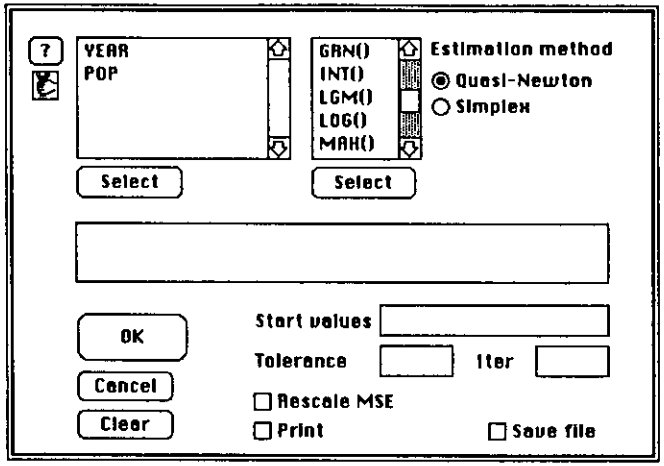
Press and hold the mouse button on **Nonlin**. A submenu pops up offering **Loss function...**, **Model...**, and **Resume**. Select the item you want and release the mouse button.

Model...

Use **Model...** to set up your model.

- Select **Nonlin/Model...** from the **Stats** menu
- Use any combination of typing and selecting from the variable and function selection lists to build your model
- Click OK

© 1989, SYSTAT, Inc.



Be careful to put the arguments for SYSTAT functions such as LOG( ) inside the parentheses.

Nonlin uses models resembling those for **MGLH** and SYSTAT supplements such as LOGIT. There is one critical difference, however. The **Nonlin** model statement is a literal algebraic expression of variables and parameters. Choose any name you wish for these parameters. Any names you specify that are not variable names in the file currently open are assumed to be parameter names.

Suppose you model Y=MU. **Nonlin** assumes that Y is a variable in the file currently open. If a variable MU is in the same file, then you have not specified any parameters in this model statement. If MU is not a variable in the file, it is assumed to be a parameter.

Another example. Suppose you set up the following model for the MEDICAL data:

```
CARDIO = B0 + B1*DIVISION
```


© 1989, SYSTAT, Inc.

Since B0 and B1 are not variables, they are parameters. The following model is mathematically equivalent:

```
CARDIO = CONSTANT + BETA*DIVISION
```

Parameter names can be any names that meet the requirements for SYSTAT variable names (8 characters beginning with a letter, excluding reserved words such as "URN"). However, unlike variable names, parameter names may not have subscripts.

The following model is similar but not equivalent to the one above, because no slope parameter has been specified:

```
CARDIO = CONSTANT + DIVISION
```

This model has only one parameter, CONSTANT. Every symbol is part of the algebraic expression in **Nonlin**. Thus, the following model's estimate of BETA is the negative of the BETA returned by the example before the last.

```
CARDIO = CONSTANT - BETA*DIVISION
```

Any legal BASIC expression can be used in a model statement, including trigonometric and other functions, plus the special variables CASE, EOF, etc. See the DATA manuals for more information. The only restriction is that the dependent variable must be a variable in your file. Here is a more complicated example:

```
CARDIO = (DIVISION<5)*MU1 + (DIVISION>=5)*MU2
```

This model has two parameters (MU1 and MU2). Their values are conditional on the value of DIVISION. Notice that the remaining parts of this expression involve logical operations (DIVISION>5). SYSTAT BASIC evaluates these to 1 (TRUE) or 0 (FALSE).

Here is an example of an exponentially damped sine wave:

```
Y = A*EXP(-B*X)*SIN(C*X)
```

The parameters in this model are A, B, and C (assuming X and Y are variables in the open file).

© 1989, SYSTAT, Inc.

Estimation method: [Quasi-Newton] and [Simplex]

Choose either the [Quasi-Newton] or [Simplex] estimation method.

The Quasi-Newton method uses numeric estimates of the first and second derivatives of your loss function to seek a minimum. If these derivatives are undefined in the region in which you are seeking minimum values, the method does not work. Specifically, the first and second derivatives must exist at all points for which the algorithm wishes to compute values.

The Simplex algorithm is a direct search procedure and therefore does not have this requirement. It calculates a value for your loss function at some point, looks to see if this value is less than values elsewhere, and steps to a new point to try again. When the steps become small, iterations stop.

Simplex is generally slower than Quasi-Newton, particularly for least squares, because Simplex cannot make use of the information in the second derivatives to find how far to move its estimates at each step. It is not a bad idea to try both algorithms on the same function, however, since they involve such radically different strategies.

Initial values: [Start values]

You may specify starting values for the parameters in the [Start values] box. Separate each value with a comma, e.g. "5,.2" for two parameters or "1E–5,123.4,15 " for three parameters You must specify the same number of values as there are parameters in your model, and the values must be in the same order as the parameters. The default values are .1 for each parameter.

Iterations

The default is 20 iterations; you can specify your own number with the [Iter] option. If you do not reach convergence in the number of iterations you specify, you can select **Resume...** to continue iterating. For some models with many parameters, you may want to begin with just a few iterations to see if all looks well and then continue.

Convergence: [Tolerance]

You may specify a [Tolerance] criterion for convergence. The default is .00005 (five hundred-thousandths of the absolute magnitude of each parameter). With this value, iterations will usually terminate when the displayed values of the parameter estimates in the iteration log fail to change. Do not confuse the word "tolerance" here with that in **MGLH**, where it refers to multicollinearity.

© 1989, SYSTAT, Inc.

You may specify any value between 1E–15 and 1. A greater tolerance criterion, say .1, gives faster, sloppier convergence; a slower criterion, say .000001, is slower and more precise. You'd use a criterion like 1E–9 only if you are compulsive.

Maximum likelihood estimation: [Rescale MSE]

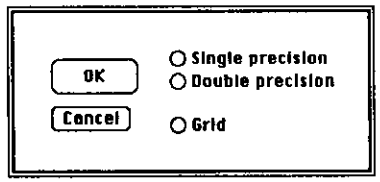
For maximum likelihood estimation, you should rescale the mean square error to 1 at the end of the iterations to get correct standard errors of the parameters. To do this, click [Rescale MSE]. The normal option (for least squares) is not to rescale.

Printing standard errors

To force **Nonlin** to print parameter estimates and standard errors even when an estimation fails to converge, click [Print].

Saving files

Click [Save file] in the **Model...** dialog box to save a file of residuals, estimated values, and variables from your model statement. You can instead save a grid of loss function values surrounding the converged minimum by clicking [Grid].



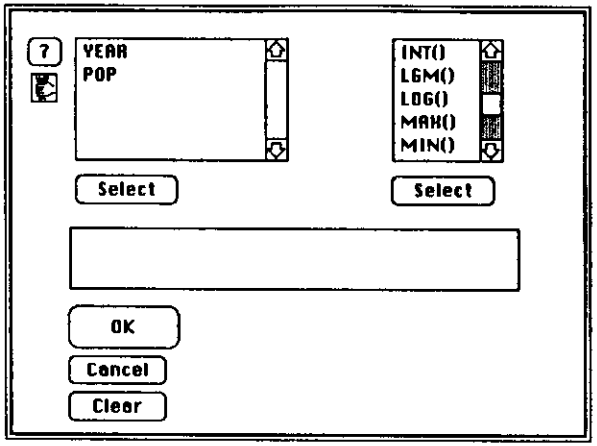
The [Grid] option takes a lot of disk space and computer time. For each pair of parameters, the file contains 784 cases (28×28), with one column for each parameter. Three parameters generate a file of 2352 cases (3 pairs times 784 cases) with three columns, for a total of 7056 data points.

Loss function...

The **Loss function...** menu allows you to compute loss functions other than least squares. Use the **Loss function...** dialog box the same way as the **Model...** dialog: use any combination of typing and selecting you need to build a function in the box. Be careful to place the arguments for SYSTAT functions—like ABS( )—inside the parentheses.

- Select **Nonlin/Loss function...** from the **Stats** menu
- Use any combination of typing and selecting from the variable and function selection lists to build your function
- Click OK

© 1989, SYSTAT, Inc.



Here is a simple loss function:

```
ABS(X-ESTIMATE)
```

This loss function is least absolute values. The word ESTIMATE in the statement is the fitted value from your model. Otherwise, you can use new or old parameters and variables in your loss statement.

Here is what least squares (the default if you do not use **Loss function...**) looks like:

```
(X-ESTIMATE)^2
```

or

```
(X-ESTIMATE)*(X-ESTIMATE)
```

The loss statement is most useful for computing maximum likelihood estimates. This is done by expressing the negative of the log-likelihood function in the loss statement, e.g.

```
-COUNT*(FAILURE*LOG(ESTIMATE) +
  (1-FAILURE)*LOG(1-ESTIMATE))
```


© 1989, SYSTAT, Inc.

Resume...

If you do not get convergence in the number of iterations allowed by your model, you can select **Nonlin/Resume** to get more iterations. **Resume** picks up where **Model...** left off.

Show complete results

If you choose [Show extended results] from **Data/Formats...**, **Nonlin** displays standard errors.

Operation

The estimation method works as follows. First, the data are read into memory, selecting only variables from the data set that were specified in the model and loss statements. Then starting values of the parameters are selected by you or the program. Then the model statement is evaluated for the first case using SYSTAT BASIC in double precision. The result of this function is called ESTIMATE. Then the loss statement is evaluated for the first case, using the estimate from the model statement. If you did not include a loss statement, then loss is computed by squaring the residual for the first case.

This procedure is repeated for all cases in the file and the loss is summed over all cases. The summed loss is then minimized by Quasi-Newton or Simplex. Iterations continue until the tolerance criterion or maximum iteration limit is reached.

Asymptotic standard errors and parameter intercorrelations

Nonlin computes asymptotic standard errors and correlations by estimating the Hessian (second derivative) matrix after iterations have terminated. Some nonlinear programs compute these statistics from an iteratively updated Hessian matrix. These estimates are known to be sensitive to iteration history and tolerance values (Scales, 1985; Dennis & Schnabel, 1983). Two different runs of these programs that converge on the same parameter values can show different standard errors and correlations. This will not happen with **Nonlin**. If the parameter estimates are the same, the standard errors will be the same.

There is a tradeoff to this increased accuracy. The Hessian estimation in SYSTAT can take as much computer time as the estimation of the parameters. Consequently, you must choose [Show extended results] from **Data/Formats...** and sit patiently if you want standard errors. Most of the examples in this chapter show the extended results.

© 1989, SYSTAT, Inc.

Occasionally, you will see the message:

```
Singular Hessian. Standard Errors not computable.
```

In these cases, the estimate of the Hessian matrix is not positive definite (all positive, nonzero roots). This warning means that your parameters are highly intercorrelated. You should examine your model for redundant parameters and possibly rescale your data so that variables are on comparable scales.

Even when you have successfully estimated your model (convergence achieved), this message is a warning to you that your model probably could be improved by dropping one or more parameters or rescaling the data. A model with fewer parameters may not fit your sample data as well, but in most cases, the simpler model generalizes better to new data. Bates and Watts (1988) give many examples showing why asymptotic standard errors can be extremely misleading. They are worse when the contours of the loss surface near the parameter minimum are not well approximated by elipsoids.

Parameter bounds and model constraints

You can put constraints on models or loss functions. We do this in Example 7.5, "Piecewise regression." Here are some more example model statements:

```
Y = (ABS(R)<=3.14159)*SIN(R/BETA)
```

In this case, Y equals 0 if the absolute value of R is greater than π. Here is a function constrained twice:

```
Y = (X<=0)*10 + (X>0 AND X<1)*BETA*X + (X>=1)*20
```

In this model, Y is 10 if X is less than or equal to 0, Y is BETA*X if X is greater than 0 and less than 1, and Y is 20 if X is greater than or equal to 1.

These types of constraints are useful for specifying bounded probability functions such as the cumulative uniform distribution.

© 1989, SYSTAT, Inc.

You can constrain your loss function as well:

```
(THETA>0 AND THETA<1)*(Y-ESTIMATE)^2 +
  (THETA<=0 OR THETA>=1)*1E25
```

In this case, if the parameter THETA falls outside of the (0,1) interval, the loss function becomes enormous. This penalty function keeps the estimates of THETA inside the interval. This method generally works with Simplex estimation. For Quasi-Newton, you may do better with more rounded corners on your bounds:

```
(Y-ESTIMATE)^2 + 1/(THETA-0)^2 + 1/(1-THETA)^2
```

You can use a **Loss...** statement to select subsets of your data:

```
(CASE<100)*(Y-ESTIMATE)^2
```

This function selects the first 99 cases because the loss for the remaining cases is 0. (Remember, CASE is a built-in variable in SYSTAT BASIC.) If you save residuals, they are computed for all cases using the parameter estimates derived from the subset. Here is how to select all cases where the variable N is 25 or 35. The variable N need not appear in your model statement.

```
(N=25 OR N=35)*(Y-ESTIMATE)^2
```

This way, you can fit different curves to different subsets of your data without leaving **Nonlin**. Be creative. Condition a loss function on your astrological sign or your loved one's favorite number.

Problems

Nonlinear estimation is an art. One-parameter estimation is minimalist, multi-parameter is rococo. There are numerous booby traps (dependencies, discontinuities, local minima, etc.) that can ruin your day. **Nonlin** offers several ways to get out of traps, but some may be hopeless. Several possibilities:

1) If you fail to get convergence with the Quasi-Newton procedure, try **Resume** to get more iterations. If this still won't work, try easing up on [Tolerance], e.g. increasing it to .1, especially if the parameter estimates seem to have converged but the program will not print the final estimates.

© 1989, SYSTAT, Inc.

2) Try switching back and forth between Quasi-Newton and Simplex without changing the starting values. That way, Simplex may help you out of a convergence or local minimum problem.

3) Use several different starting values for both methods before you feel comfortable with the final estimates. This can help you expose local minima. Usually, the farther the initial starting values are from the final ones, the better is the parameter correlation and standard error information. The Simplex method is most robust against local minima if you start it with outrageous initial values. There is a tradeoff, however: it can take forever to get back to friendly parameter territory.

4) Simplex is immune to round-off error because it computes new estimates every time it iterates. If the Quasi-Newton method fails to converge because of round-off error, you can use Simplex to get to your goal inexorably. It fails only when two separate values of the function itself differ by less than machine precision (1E-15).

5) If you get illegal function values for starting values, try some other estimates. For some functions with many parameters, you may need high quality starting values to get an estimable function at all.

6) Simplex does not require that the model be differentiable in the region you are exploring the way Quasi-Newton and most other descent methods do. Thus, Simplex may be required for some loss functions, for example least absolute values. Sometimes you can use Simplex to get into a differentiable neighborhood and switch to Quasi-Newton to finish off the estimates. Furthermore, the farther afield your starting parameter values are, the more widely Simplex looks for minima.

7) Never trust the output of an iterative nonlinear estimation procedure until you have plotted estimates against predictors and you have tried several different starting values. The [Save file] option was designed so that you can quickly save estimates, residuals, and model variables and plot them. All the examples in this chapter were tested this way. Although most began with default starting values for the parameters (.1), they were checked with other starting values.

8) The "Round-off Error" message with Quasi-Newton occurs when truncation or cancellation errors make it impossible to compute first derivatives using finite differences. If you get this message, try a new starting point with the [Start values] option.

© 1989, SYSTAT, Inc.

Computation

Missing data

Missing values are handled according to the conventions of SYSTAT BASIC. That is, missing values propagate in algebraic expressions. For example, "X + ." is a missing value. The expression "X = ." is not missing, however. It is 1 if X is missing and 0 if not. Thus, you can use logical expressions to put conditions on model or loss functions; consider the following loss function:

```
(X<>.)*(Y-ESTIMATE)^2 + (X=.)*(Z-ESTIMATE)^2
```

Illegal expressions (such as division by zero and negative square roots) are set to missing values. If this happens when computing the loss statement for a particular case, the loss function is set to an extremely large value (10^{35}). This way, parameter estimates are forced to move away from regions of the parameter space that yield illegal function evaluations.

Overflows (such as a positive number with an extremely large exponent) are set to machine overflow (10^{35}). Negative overflows are set to the negative of this value. Overflows usually cause the loss function to be large, so the program is forced to move away from estimates that produce overflows.

These features mean that **Nonlin** tends to "crash" less frequently than most other nonlinear estimation programs. It will continue for several iterations to try parameter values that lower the loss value, even when some of these lead to a seemingly hopeless result. It is your responsibility to check whether final estimates are reasonable, however, by using both estimation methods, different starting values, and other options.

Algorithms

The Quasi-Newton method is described in Fletcher (1972) and is sometimes called modified Fletcher/Powell. Modifications include the LDL' Cholesky factorization of the updated Hessian matrix. It is the same algorithm employed in **Series** for ARIMA estimation. The Simplex method is adapted from O'Neill (1971), with several revisions noted in Griffiths and Hill (1985).

© 1989, SYSTAT, Inc.

The loss function is computed in two steps: First, the model statement is evaluated for a case using current values of the parameters and data. Second, the **Loss...** statement is evaluated using ESTIMATE (computed as the result of the model statement evaluation) and other parameter and data values. These two steps are repeated for all cases, over which the result of the loss function is summed. The summed **Loss...** is then minimized by the Quasi-Newton or Simplex procedure. Step halvings are used in the minimizations when model or loss statement evaluations overflow or result in illegal values. If repeated step halvings down to machine epsilon (error limit) fail to remedy this situation, iterations cease with an "Illegal values" message.

Asymptotic standard errors are computed by the central differencing finite approximation of the Hessian matrix. Some nonlinear regression programs compute standard errors by squaring the Jacobian matrix of first derivatives. Others use different methods altogether. For linear models, all valid methods produce identical results. For some nonlinear models, however, the results may differ. The Hessian approach, which works well for nonlinear regression, is also ideally suited for **Nonlin**'s maximum likelihood estimation.

© 1989, SYSTAT, Inc.

Examples

The following examples illustrate a few of the applications of **Nonlin**. Like **MGLH**, it encompasses a variety of models and allows you to solve problems that are often "hard-wired" in other packages.

7.1 Estimating a mean

The sample mean is a least squares estimate of location. That is, the mean is the point at which the squared deviations of sample values are minimum. This is the simplest model you can fit with **Nonlin**. Here is an example with the IRIS data:

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Click OK

- Select **Select...** from the **Data** menu
- Fill in the Select statement: Select SPECIES=1
- Click OK

- Select **Nonlin/Model...** from the **Stats** menu
- Build the model statement: SEPALLEN=MEAN
- Click OK

```
Iteration       Loss       Parameter values
    0       .12095300+04  .1000D+00
    1       .60882000+01  .5006D+01
    2       .60882000+01  .5006D+01

Dependent variable is SEPALLEN

   Source   Sum-of-squares   DF  Mean-square

 Regression       1253.002    1     1253.002
   Residual          6.088   49        0.124

      Total       1259.090   50
  Corrected          6.088   49

      Raw R-squared (1-Residual/Total)     =        0.995
Corrected R-Squared (1-Residual/Corrected) =       -0.000


Parameter        Estimate       A.S.E.       Lower  <95%> Upper
    MEAN            5.006        0.050       4.906        5.106
```

Notice you get the standard error of the mean along for the ride. The 95 percent confidence interval is computed using the *t* distribution with 49 degrees of freedom.

© 1989, SYSTAT, Inc.

7.2 Estimating a median

The sample median minimizes absolute deviations instead of squared deviations. Notice that for our example, we do not choose [Show complete results]. If you tried the last example and want to try this one, too, first choose [Show standard results] from **Data/Formats...**

- Select **Select...** from the **Data** menu
- Fill in the Select statement: Select SPECIES=1
- Click OK

- Select **Nonlin/Loss...** from the **Stats** menu
- Build the loss statement: ABS(SEPALLEN–ESTIMATE)
- Click OK

- Select **Nonlin/Model...** from the **Stats** menu
- Build the model statement: SEPALLEN=MEDIAN
- Choose [Simplex] estimation
- Click OK

We use Simplex estimation because the first and second derivatives of the loss function do not exist at the median. For the same reason, we have not chosen [Show extended results], because the standard errors cannot be estimated at the exact minimum. Notice that the program does not print an analysis of variance summary table either. It will not print this table if you use the loss statement, because it cannot be sure you are using least squares (the default).

Note that we do not need to sort to get the estimate. For large samples, sorting time is offset by the time it takes to evaluate and minimize the loss statement, however. You might want to try both methods.

```
Iteration       Loss       Parameter values
    0       .24530000+03  .10000+00
    1       .14500000+02  .50500+01
    2       .13500590+02  .50000+01
    3       .13500000+02  .50000+01

Dependent variable is SEPALLEN
Final value of loss function is        13.500


Parameter        Estimate
  MEDIAN            5.000
```

For samples with an even number of cases, any value between the two middle observations is a solution to the problem. The conventional method of averaging the two middle sorted cases is based on a different definition. See Hoaglin, Mosteller, and Tukey (1983) and Maindonald (1984) for more details.

© 1989, SYSTAT, Inc.

You can use **Nonlin** to compute other weighted estimates of location, such as the biweight. To do some of these, you need first to sort the file and use the built in variable CASE to compute biweights.

7.3 Linear regression

You would not want to use **Nonlin** for linear regression since **MGLH** is available. Nevertheless, to illustrate how **Nonlin** works with more than one parameter, let's redo the **MGLH** chapter's Example 6.1, "Simple regression" from Metzler and Shepard (1974). The example uses the file ROTATE.

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Click OK

- Select **Nonlin/Model...** from the **Stats** menu
- Build the statement: RT = CONSTANT + BETA*ANGLE
- Click OK

Notice the difference between the model statement here and the one in MGLH. You must specify the slope parameter (BETA) explicitly. You may use any names you wish for the constant and slope parameters, but they may not be names of variables in the file. **MGLH** does not require the slope parameter, because all of its models are linear.

```
Iteration       Loss       Parameter values
    0       .51170000+03  .10000+00 .10000+00
    1       .12701070+03  .99310-01 .35690-02
    2       .60064680+02  .51900+01-.20380-01
    3       .19603330+01  .16840+01 .25590-01
    4       .76723410+00  .19150+01 .20830-01
    5       .76722580+00  .19160+01 .20820-01
    6       .76722580+00  .19160+01 .20820-01

Dependent variable is       RT

   Source   Sum-of-squares    DF  Mean-square

 Regression       161.873     2       80.936
   Residual         0.767     8        0.096

      Total       162.640    10
  Corrected        15.184     9

      Raw R-squared (1-Residual/Total)     =        0.995

Corrected R-Squared (1-Residual/Corrected) =        0.949


Parameter        Estimate       A.S.E.       Lower  <95%> Upper
CONSTANT            1.916        0.185       1.489        2.342
    BETA            0.021        0.002       0.017        0.025
```

© 1989, SYSTAT, Inc.

```
Asymptotic Correlation Matrix of Parameters

                CONSTANT        BETA

CONSTANT           1.000
    BETA          -0.848       1.000
```

7.4 Least absolute values regression

We can estimate the same function using least absolute values loss instead of least squares. Just add a **Loss function...**:

- Select **Nonlin/Loss function...** from the **Stats** menu
- Build the statement: ABS(RT–ESTIMATE)

- Select **Nonlin/Model...** from the **Stats** menu
- Build the statement: RT = CONSTANT + BETA*ANGLE
- Choose [Simplex] estimation
- Click OK

ESTIMATE is a keyword here in the loss statement. We choose [Simplex] estimation because the loss function has no derivative at its minimum.

```
Iteration       Loss       Parameter values
    0       .58200000D+02  .10000D+00 .10000D+00
    1       .86625000D+01  .28440D+00 .31250D-01
    2       .5261328D+01   .12600D+01 .2725D-01
    3       .2222124D+01   .19480D+01 .2064D-01
    4       .21539400D+01  .19400D+01 .20890-01
    5       .2150894D+01   .19400D+01 .2089D-01
    6       .21506450+01   .19410D+01 .2089D-01

Dependent variable is       RT
Final value of loss function is        2.151


Parameter        Estimate
CONSTANT            1.941
    BETA            0.021
```

In this case, the function is almost identical to that estimated by least squares. That is because there are no serious outliers in the Metzler and Shepard data.

7.5 Nonlinear regression

The following data (USPOP) comprise the U.S. population (in thousands) beginning in 1790.

© 1989, SYSTAT, Inc.

YEAR	POP
1790	3929
1800	5308
1810	7239
1820	9638
1830	12866
1840	17069
1850	23191
1860	31443
1870	39818
1880	50155
1890	62947
1900	75994
1910	91972
1920	105710
1930	122775
1940	131669
1950	151325
1960	179323
1970	203211

We fit an exponential model to these data:

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Click OK

- Select **Nonlin/Model...** from the **Stats** menu
- Build the statement: POP = EXP(C + B1*YEAR)
- Specify 0,0 for [Start values]
- Click [Save file]
- Click OK

The starting values are required, since default starting values of .1 for these parameters result in the function overflowing to an illegal value.

```
Iteration       Loss       Parameter values
    0       .16440260+12   .00000+00 .00000+00
    1       .1445300D+12   .24050-05 .46450-02
    2       .48762250+11   .29620-05 .57230-02
    3       .31708630+11   .31440-05 .60810-02
    4       .22258000+11  -.61560+01 .93410-02
    5       .41890060+10  -.1177D+02 .12130-01
    6       .13783080+10  -.1850D+02 .15630-01
    7       .1144494D+10  -.1853D+02 .15630-01
    8       .10905900+10  -.19160+02 .15950-01
    9       .10873340+10  -.1939D+02 .16070-01
   10       .10872510+10  -.1942D+02 .16080-01
   11       .10872490+10  -.1942D+02 .16090-01
   12       .10872490+10  -.1942D+02 .16090-01
   13       .10872450+10  -.1943D+02 .16090-01
   14       .10872450+10  -.1943D+02 .16090-01
```

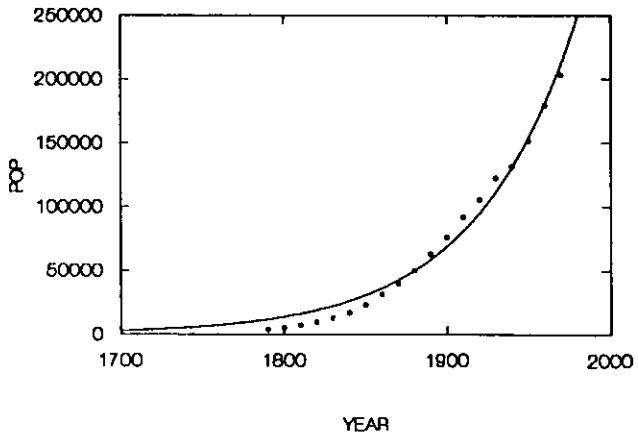
© 1989, SYSTAT, Inc.

```
Dependent variable is      POP

   Source   Sum-of-squares    Df  Mean-square

 Regression     .163318E+12     2  .816590E+11
   Residual     .108724E+10    17  .639556E+08

      Total     .164405E+12    19
  Corrected     .719228E+11    18

     Raw R-squared (1-Residual/Total)     =        0.993
Corrected R-Squared (1-Residual/Corrected) =        0.985

Singular Hessian, Standard errors not computable.

Parameter        Estimate
        C         -19.434
       B1           0.016
```

Let's look at a plot of the function. First, open the file containing the residuals. We have added the predicted line with **Function.**

- Select **Plot/Plot** from the **Graph** menu
- Select the Y variable: POP
- Select the X variable: YEAR
- Click OK

- Select **Function** from the **Graph** menu
- Build the equation: Y=EXP(–19.434+.016*X)
- Set [X min] to 1700, [X max] to 2000, [Y min] to 0, [Y max] to 250000, [X label] to " ", [Y label] to " " in the [Axes] dialog box
- Click OK

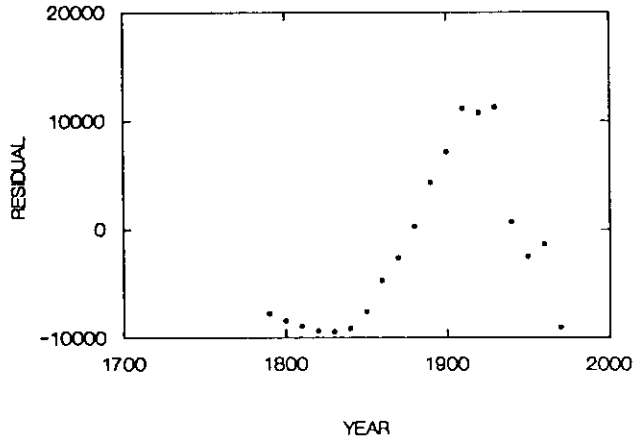


© 1989, SYSTAT, Inc.

The growth curve looks OK only in portions of the range.

We can check the residuals by plotting RESIDUAL against YEAR.

- Select **Plot/Plot** from the **Graph** menu
- Select the Y variable: RESIDUAL
- Select the X variable: YEAR
- Click OK



Not surprisingly, there appears to be some serial correlation in the residuals. We can check the serial correlation in the residuals further by using **Series**.

- Select **Series/ACF plot...** from the **Stats** menu
- Select the variable: RESIDUAL
- Click OK

```
Residual copied from SYSTAT file into active work area

Plot of RESIDUAL
Number of cases =   19
Mean of series =   -1850.241
Standard deviation of series =    7334.849
```

© 1989, SYSTAT, Inc.

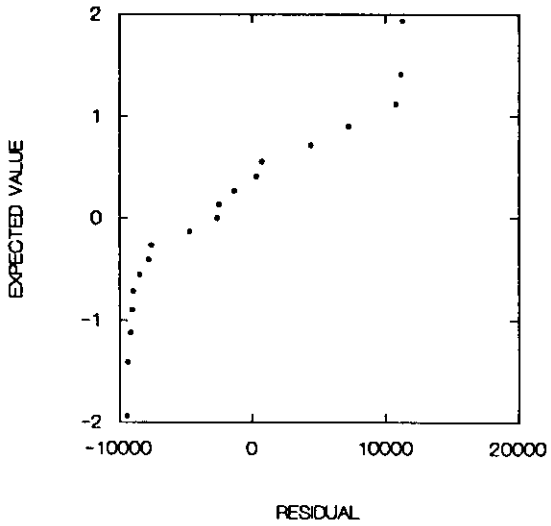
Plot of autocorrelations

Lag	Corr	SE	-1.0 -.8 -.6 -.4 -.2 .0 .2 .4 .6 .8 1.0
1	.837	.229	
2	.658	.355	
3	.409	.415	
4	.107	.435	
5	-.097	.437	
6	-.273	.438	
7	-.370	.447	
8	-.428	.463	
9	-.426	.483	
10	-.385	.502	
11	-.318	.518	
12	-.225	.528	
13	-.124	.533	
14	-.039	.534	
15	.039	.534	

The first autocorrelation looks significant and there are additional periodicities in the residuals.

Let's look at a normal probability plot:

- Select **Prob'y/Normal...** from the **Graph** menu
- Select the variable: RESIDUAL
- Click OK



© 1989, SYSTAT, Inc.

This plot shows the residuals to be positively skewed. You might want to try a histogram or stem and leaf diagram on them.

Notice how useful other SYSTAT and SYGRAPH items can be for diagnosing the results of your nonlinear estimation. You should use the graphics frequently for assessing the adequacy of a model. You can use SYGRAPH to plot smooth functions that you have estimated, as we did in this example. Alternatively, the [Smooth] option in **Plot** allows you to draw a [Spline] curve through the ESTIMATE values in your SYSTAT file created by **Nonlin**.

7.6 Piecewise regression

Sometimes we need to fit two different regression functions to the same data. For example, sales of a certain product might be strongly related to quality when advertising budgets are below a certain level—that is, when sales are generated by "word of mouth." Above this ad budget level, sales may be less strongly related to quality of goods and more by marketing and advertising factors. In these cases, we can fit different sections of the data with different models. It is easier to combine these into a single model, however.

Here is an example of a quadratic function with a ceiling using data from Gilfoil (1982). This particular study is one of several that show that dialog menu interfaces are preferred by inexperienced computer users, and command based interfaces are preferred by experienced users. (This is why we provide both a standard Macintosh interface and a command interface.) The data below are for one subject. The variable SESSION is the session number and TASKS is the number of user-controlled tasks (as opposed to dialog) chosen by the subject during a session. The data are in the file LEARN.

© 1989, SYSTAT, Inc.

SESSION	TASKS
1	0
2	0
3	0
4	1
5	0
6	1
7	1
8	6
9	6
10	6
11	5
12	6
13	6
14	6
15	6
16	6
17	6
18	6
19	6
20	6

We will fit these data with a quadratic model for earlier sessions and a ceiling for later sessions. We use **Nonlin** to estimate the point where the learning hits this ceiling (at 6 tasks).

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Click OK

- Select **Nonlin/Model...** from the **Stats** menu
- Build the statement: TASKS=(B*SESSION^2)*(SESSION<KNOW) + (B*KNOW^2)*(SESSION>=KNOW)
- Specify .02, 8 for [Starting values]
- Click OK

Note that the expressions (SESSION<KNOW and SESSION>=KNOW) control which function is to be used: the quadratic or the horizontal line. Here is the resulting output.

© 1989, SYSTAT, Inc.


```
Iteration       Loss       Parameter values
    0      .28200960D+03  .2000D-01 .8000D+01
    1      .10936750D+03  .1253D+00 .8000D+01
    2      .97813300D+02  .1584D+00 .4840D+01
    3      .61783050D+02  .1410D+00 .5851D+01
    4      .42431610D+02  .1146D+00 .7240D+01
    5      .15036740D+02  .6654D-01 .9586D+01
    6      .14467580D+02  .6208D-01 .9796D+01
    7      .14429590D+02  .6250D-01 .9731D+01
    8      .14418180D+02  .6336D-01 .9656D+01
    9      .14418120D+02  .6333D-01 .9660D+01
   10      .14418120D+02  .6333D-01 .9660D+01

Dependent variable is    TASKS

   Source   Sum-of-squares    Df  Mean-square

 Regression        445.582     2      222.791
   Residual         14.418    18        0.801

      Total        460.000    20
  Corrected        140.000    19

      Raw R-squared (1-Residual/Total)     =        0.969
Corrected R-Squared (1-Residual/Corrected) =        0.897

Parameter        Estimate        A.S.E.        Lower  <95%> Upper
        B           0.063         0.007        0.048        0.079
     KNOW           9.660         0.594        8.412       10.907

Asymptotic Correlation Matrix of  Parameters

                        B        KNOW

        B           1.000
     KNOW          -0.928       1.000
```

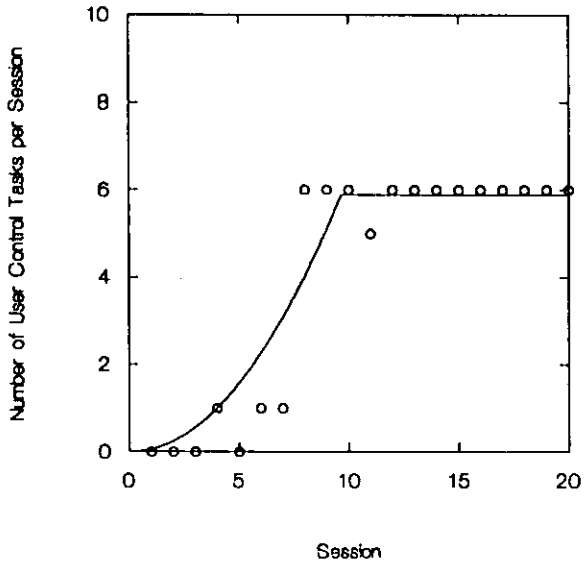
We can use the model equation with **Function** to plot these data and the fitted curve.

© 1989, SYSTAT, Inc.

- Select **Function** from the **Graph** menu
- Build the equation: TASKS=(.063*SESSION^2)*(SESSION<9.66) + (.063*9.66^2)*(SESSION=>9.66)
- Set [X min] to 0, [X max] to 20, [Y min] to 0, [Y max] to 10, [X label] to "Session", [Y label] to "Number of User Control Tasks per Session" in the [Axes] dialog box
- Click OK

- Select **Open...** from the **File** menu
- Select the LEARN file
- Click Open

- Select **Plot/Plot** from the **Graph** menu
- Select the Y variable: TASKS
- Select the X variable: SESSION
- Set [X min] to 0, [X max] to 20, [Y min] to 0, [Y max] to 10, [X label] to " ", [Y label] to " "
- Click OK



© 1989, SYSTAT, Inc.

The fit at the lower end is not impressive. We might want to fit a truncated logistic model (see below in this chapter) instead of a quadratic because learning is more often represented with this type of function. This model would have a logistic curve at the lower values of SESSION and a flat ceiling line at the upper end. We should use a **Loss function...** also to make the fit maximum likelihood.

Piecewise linear regression models with unknown breakpoints can be fitted similarly. These models look like this:

```
Y = B0 + B1*X + B2*(X-BREAK)*(X>BREAK)
```

If the break point is known, then you can use **MGLH** to do ordinary regression to fit the separate pieces. See Neter, Wasserman, and Kutner (1985) for an example.

7.7 Maximum likelihood estimation

Because **Nonlin** includes a loss function, you can maximize the likelihood of a function in the model equation. The way to do this is to minimize the negative of the log-likelihood.

Here is an example. Let's compute the maximum likelihood estimates of the mean and variance of SEPALLEN assuming a normal distribution for the first species in the IRIS data set. For a sample of N independent normal random variables, the log-likelihood function is:

$$L(\mu,\sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(X-\mu)^2$$

We minimize the negative of this log-likelihood with **Loss function...** and thus maximize the likelihood.

© 1989, SYSTAT, Inc.

- Select **Select...** from the **Data** menu
- Fill in the Select statement: Select SPECIES=1
- Click OK

- Select **Nonlin/Loss function...** from the **Stats** menu
- Build the statement: .5*LOG(2*3.14159) + .5*LOG(VARIANCE) + (1/(2*VARIANCE))*(SEPALLEN-ESTIMATE)^2
- Click OK

- Select **Nonlin/Model...** from the **Stats** menu
- Build the statement: SEPALLEN=MEAN
- Choose [Simplex] estimation
- Click [Rescale MSE]
- Click OK

Notice that since the loss is summed over all cases, we do not use *N* (50 in this example) in the loss sentence. Notice, too, that VARIANCE does not appear in the model. It is an extra parameter introduced in the computation of the loss function.

We are using the [Simplex] option to save space, since it converges in fewer iterations in this case.

```
Iteration       Loss       Parameter values
    0       .6036032D+04  .1000D+00 .1000D+00
    1       .1488101D+03  .3651D+01 .1982D+02
    2       .9530052D+02  .4466D+01 .5538D+00
    3       .3325912D+02  .4995D+01 .3112D+00
    4       .1835365D+02  .5006D+01 .1184D+00
    5       .1830520D+02  .5006D+01 .1217D+00
    6       .1830514D+02  .5006D+01 .1218D+00
    7       .1830514D+02  .5006D+01 .1218D+00

Dependent variable is SEPALLEN
Final value of loss function is       18.305

Parameter       Estimate

     MEAN           5.006
 VARIANCE           0.122
```

Notice that the least squares estimate of the variance in the Stats chapter is larger than the biased maximum likelihood estimate here.

© 1989, SYSTAT, Inc.

7.8 Logistic models—iteratively reweighted least squares

Cox (1970) reports the following data on tests among objects for failures after certain times. FAILURE is the number of failures and COUNT is the total number of tests in the file COX.

TIME	FAILURE	COUNT
1	0	55
14	2	157
27	1	159
51	3	16

Cox uses a logistic model for fitting these results:

```
ESTIMATE = COUNT*EXP(B0 + B1*TIME)/
  (1+EXP(B0 + B1*TIME))
```

Because the counts differ at each time, the variances of the failures also differ. If FAILURE is randomly sampled from a binomial, then

```
VAR(FAILURE) = ESTIMATE*(COUNT - ESTIMATE)/COUNT
```

We use these variances to weight each case in the estimation. On each iteration, the variances are recalculated from the new estimates and used anew in computing the weighted loss function. Here is how to do it:

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Click OK

- Select **Nonlin/Loss function...** from the **Stats** menu
- Build the statement: (FAILURE – ESTIMATE)^2*COUNT /(ESTIMATE*(COUNT – ESTIMATE))
- Click OK

- Select **Nonlin/Model...** from the **Stats** menu
- Build the statement FAILURE = COUNT*EXP(B0 + B1*TIME)/(1+EXP(B0 + B1*TIME))
- Click [Rescale MSE]
- Click OK

Notice that the first part of the loss function (FAILURE – ESTIMATE)^2 is just least squares and the second part is the weight to be applied to a given case (namely the reciprocal of the variance).

© 1989, SYSTAT, Inc.

```
Iteration       Loss       Parameter values
    0       .50902670+04   .10000+00 .10000+00
    1       .12818250+03   .9512D-01 -.64630-01
    2       .88282440+02   -.1027D+02 .23740+00
    3       .10633570+02   -.7175D+01 .14540+00
    4       .31830280+01   -.5864D+01 .10640+00
    5       .22713810+01   -.5334D+01 .90330-01
    6       .21043370+01   -.5027D+01 .80440-01
    7       .12946080+01   -.5344D+01 .86360-01
    8       .70888420+00   -.5417D+01 .82870-01
    9       .65619350+00   -.5343D+01 .79130-01
   10       .65427610+00   -.5314D+01 .78160-01
   11       .65425970+00   -.5311D+01 .78090-01
   12       .65425960+00   -.5311D+01 .78090-01

Dependent variable is  FAILURE
Final value of loss function is        0.654

Parameter        Estimate       A.S.E.        Lower  <95%> Upper
      B0             -5.311        0.434       -7.179       -3.442
      B1              0.078        0.014        0.020        0.136

Asymptotic correlation matrix of parameters

                        B0          B1

        B0           1.000
        B1          -0.921       1.000
```

Jennrich and Moore (1975) show that this method can be used for maximum likelihood estimation of parameters from a distribution in the exponential family. There is a more direct way to get maximum likelihood estimates with **Nonlin**, however, by specifying the likelihood function directly in the loss statement; see the next example.

7.9 Logistic models—maximum likelihood estimation

This example is a reanalysis of these data in a slightly different form (COX2).

TIME	FAILURE	COUNT
7	0	55
14	0	155
14	1	2
27	0	152
27	1	7
51	0	13
51	1	3

Notice that FAILURE now is binary and COUNT is the count of failures (or successes). The log-likelihood function for the logit model is:

$$L(\beta_0, \beta_1) = \sum[p\ln(\text{ESTIMATE}) + (1-p)\ln(1-\text{ESTIMATE})]$$

where

© 1989, SYSTAT, Inc.

$$\text{ESTIMATE} = \frac{\text{EXP(B0+B1*TIME)}}{1 + \text{EXP(B0+B1*TIME)}}$$

and the sum is over all observations.

Now we set up the model and specify negative log-likelihood as the loss function.

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Choose 6 [Decimals to show]
- Click OK

- Select **Nonlin/Loss function...** from the **Stats** menu
- Build the statement: -COUNT*(FAILURE*LOG(ESTIMATE) + (1-FAILURE)*LOG(1-ESTIMATE))
- Click OK

- Select **Nonlin/Model...** from the **Stats** menu
- Build the statement: FAILURE = EXP(B0+B1*TIME)/(1 + EXP(B0+B1*TIME))
- Click [Rescale MSE]
- Click OK

We are using the [Rescale MSE] option to rescale the Mean Square Error to 1 when computing the asymptotic standard errors of the parameters.

© 1989, SYSTAT, Inc.

```
Iteration       Loss       Parameter values
    0      .8312067D+03  .10000+00 .10000+00
    1      .1194279D+03  .80460-01-.2998D+00
    2      .6382913D+02 -.1053D+02 .2182D+00
    3      .6214210D+02 -.8023D+01 .9673D-01
    4      .5688932D+02 -.8715D+01 .1413D+00
    5      .5562096D+02 -.8903D+01 .1696D+00
    6      .5316296D+02 -.8036D+01 .1355D+00
    7      .4821113D+02 -.5354D+01 .6906D-01
    8      .4779995D+02 -.5134D+01 .7019D-01
    9      .4771120D+02 -.5544D+01 .8547D-01
   10      .4768736D+02 -.5415D+01 .8057D-01
   11      .4768728D+02 -.5415D+01 .8069D-01
   12      .4768728D+02 -.5415D+01 .8070D-01
   13      .4768728D+02 -.5415D+01 .8070D-01

Dependent variable is  FAILURE
Final value of loss function is    47.687279
Standard errors of parameters are rescaled


Parameter        Estimate       A.S.E.        Lower  <95%> Upper
       B0       -5.415177     0.727585    -7.195513    -3.634842
       B1        0.080696     0.022357     0.025991     0.135401


Asymptotic correlation matrix of parameters
                       B0          B1

        B0          1.000000
        B1         -0.910147    1.000000
```

7.10 Probit models—maximum likelihood estimation

The probit model closely resembles the logit model, with the normal cumulative distribution function replacing the cumulative logistic as the response function. This substitution is easy to do in **Nonlin**, since the ZCF function for the normal distribution is available.

The following data were reported by Haberman (1982). They represent the number of men at different levels of education who believe that women should stay at home rather than work. The variable RESPONSE is 0 for those who believe women should stay at home and 1 otherwise. The data are in the file HABERMAN.

© 1989, SYSTAT, Inc.

RESPONSE	YEAR	COUNT
0	0	4
0	1	2
0	2	4
0	3	6
0	4	5
0	5	13
0	6	25
0	7	27
0	8	75
0	9	29
0	10	32
0	11	36
0	12	115
0	13	31
0	14	28
0	15	9
0	16	15
0	17	3
0	18	1
0	19	2
0	20	3
1	0	2
1	1	0
1	2	0
1	3	3
1	4	5
1	5	7
1	6	9
1	7	15
1	8	49
1	9	29
1	10	45
1	11	59
1	12	245
1	13	70
1	14	79
1	15	23
1	16	110
1	17	29
1	18	28
1	19	13
1	20	20

© 1989, SYSTAT, Inc.

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Click OK

- Select **Nonlin/Loss function...** from the **Stats** menu
- Build the statement: -COUNT*(RESPONSE*LOG(ESTIMATE) + (1-RESPONSE)*LOG(1-ESTIMATE))
- Click OK

- Select **Nonlin/Model...** from the **Stats** menu
- Build the statement: RESPONSE = ZCF(B0+B1*YEAR)
- Click [Rescale MSE]
- Click [Save file]
- Click OK

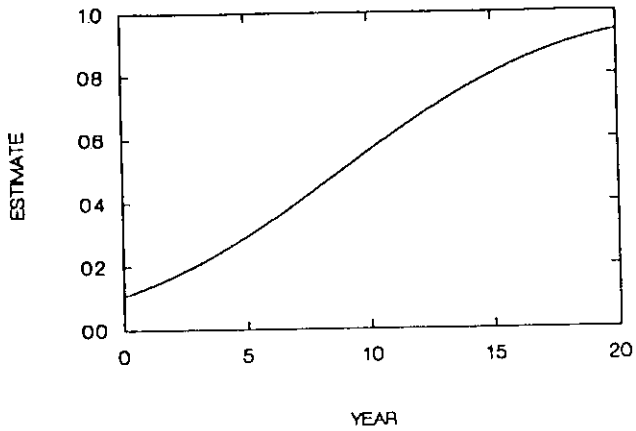
Notice we use the same loss function as in the previous logit example. This is because both functions have the same loss function when expressed in terms of the variable ESTIMATE.

```
ITERATION       LOSS          PARAMETER VALUES
    0       0.10520670+04 0.1000D+000.10000+00
    1       0.8606654D+03 0.9623D-010.59870-01
    2       0.8504258D+03 -.2520D+010.27490+00
    3       0.7688038D+03 -.1239D+010.13570+00
    4       0.7681568D+03 -.1265D+010.14180+00
    5       0.7681386D+03 -.1246D+010.1399D+00
    6       0.7681386D+03 -.1247D+010.1399D+00
    7       0.7681386D+03 -.1247D+010.1399D+00

DEPENDENT VARIABLE IS RESPONSE
FINAL VALUE OF LOSS FUNCTION IS      768.139
STANDARD ERRORS OF PARAMETERS ARE RESCALED


PARAMETER        ESTIMATE       A.S.E.       LOWER  <95%> UPPER
       B0          -1.247        0.137      -1.523       -0.971
       B1           0.140        0.012       0.117        0.163

ASYMPTOTIC CORRELATION MATRIX OF PARAMETERS

                       B0          B1

       B0           1.000
       B1          -0.962       1.000
```

Now we can plot the fitted values with SYGRAPH.

- Select **Function** from the **Graph** menu
- Build the equation: ESTIMATE=ZCF(-1.247+.14*YEAR)
- Set [X min] to 0, [X max] to 20, [Y min] to 0, [Y max] to 1
- Click OK

© 1989, SYSTAT, Inc.



Note the ogive shape of the fitted function, which is bounded by 0 and 1.

7.11 Kinetic models

You can also use **Nonlin** to test kinetic models. The following analysis models competitive inhibition for an enzyme inhibitor. The data are adapted from a conference session on statistical computing with microcomputers (Greco *et al.*,1982). We will fit three variables: initial enzyme velocity (V), concentration of the substrate (S), and concentration of the inhibitor (I). The parameters of the model are the maximum velocity (VMAX), the Michaelis constant (KM), and the dissociation constant of the enzyme-inhibitor complex (KIS). The model to estimate is:

$$V = VMAX\frac{S}{KM(1+\frac{I}{KIS}) + S}$$

The data are in the file ENZYME.

© 1989, SYSTAT, Inc.

V	S	I
.021	.25	.5
.022	.25	.5
.038	.50	.5
.036	.50	.5
.081	1.25	.5
.085	1.25	.5
.162	2.50	.5
.150	2.50	.5
.225	3.75	.5
.210	3.75	.5
.294	5.00	.5
.269	5.00	.5
.048	.25	.2
.048	.25	.2
.088	.50	.2
.089	.50	.2
.188	1.25	.2
.177	1.25	.2
.318	2.50	.2
.318	2.50	.2
.447	3.75	.2
.447	3.75	.2
.553	5.00	.2
.545	5.00	.2
.545	5.00	.2
.303	.25	.0
.310	.25	.0
.451	.50	.0
.465	.50	.0
.752	1.25	.0
.694	1.25	.0
.950	2.50	.0
.929	2.50	.0
.020	3.75	.0
.013	3.75	.0
.072	5.00	.0
.059	5.00	.0
.323	.25	.0
.310	.25	.0
.479	.50	.0
.454	.50	.0
.756	1.25	.0
.723	1.25	.0
.964	2.50	.0
.054	3.75	.0
.040	3.75	.0
.094	5.00	.0

© 1989, SYSTAT, Inc.

Here is how to do the analysis:

- Select **Formats...** from the **Data** menu
- Choose [Show extended results]
- Click OK

- Select **Nonlin/Model...** from the **Stats** menu
- Build the statement: V=VMAX*S/(KM*(1+I/KIS)+S)
- Click OK

```
Iteration       Loss       Parameter values
    0      .11884920+02  .10000+00 .10000+00 .10000+00
    1      .93481720+01  .90530+00 .44800-01 .11000+00
    2      .43012900+01  .1224D+01 .27500+00 .51170-01
    3      .14825220+01  .13030+01 .32700+00 .13810-01
    4      .13954260+01  .13130+01 .33210+00 .91170-02
    5      .54550610+00  .10600+01 .25510+00 .11380-01
    6      .47663730+00  .98840+00 .23840+00 .10250-01
    7      .38941700+00  .10300+01 .29020+00 .9594D-02
    8      .17383310+00  .1121D+01 .44560+00 .13830-01
    9      .34559020-01  .12050+01 .6765D+00 .21310-01
   10      .14386330-01  .12500+01 .8146D+00 .2597D-01
   11      .13764800-01  .12570+01 .8383D+00 .2678D-01
   12      .13703270-01  .12600+01 .8477D+00 .2711D-01
   13      .13701490-01  .12600+01 .8488D+00 .2717D-01
   14      .13698130-01  .12600+01 .84750+00 .2719D-01
   15      .13697860-01  .12600+01 .84690+00 .2718D-01
   16      .13697850-01  .12600+01 .8469D+00 .2718D-01
   17      .13697850-01  .12600+01 .84690+00 .2718D-01

Dependent variable is        V

   Source   Sum-of-squares    DF  Mean-square

 Regression      15.404014     3     5.134671
   Residual       0.013698    43     0.000319

      Total      15.417712    46
  Corrected       5.763071    45

      Raw R-squared (1-Residual/Total)     =      0.999112
Corrected R-Squared (1-Residual/Corrected) =      0.997623

Parameter        Estimate       A.S.E.       Lower  <95%> Upper
     VMAX        1.259516     0.012209    1.234894     1.284138
       KM        0.846855     0.026935    0.792536     0.901175
      KIS        0.027176     0.000878    0.025405     0.028946

Asymptotic Correlation Matrix of  Parameters

                VMAX          KM          KIS

      VMAX      1.000000
        KM      0.869400    1.000000
       KIS      0.479753    0.682255    1.000000
```

© 1989, SYSTAT, Inc.

You could try alternative models for these data such as one for uncompetitive inhibition:

$$V=VMAX*S/[KM+S*(1+I/KII))$$

or one for noncompetitive inhibition:

$$V=VMAX*S/(KM*(1+I/KIS)+S*(1+I/KII))$$

where KII is the dissociation constant of the enzyme-inhibitor-substrate complex.

7.12 Deterministic models—minimizing an analytic function

You can also use **Nonlin** to find the minimum of an algebraic function. Since this requires no data, you need a trick. Make the model a dummy statement and the loss statement the function you wish to minimize. First, make a data set containing one observation:

- Select **New** from the **File** menu
- Name the first variable: DUMMY
- Enter one value for DUMMY: 0
- Select **Save** from the **File** menu

Now do this, using the dummy file:

- Select **Nonlin/Loss function...** from the **Stats** menu
- Build the statement: 100*(U–V^2)^2 + (1–V)^2
- Click OK

- Select **Nonlin/Model...** from the **Stats** menu
- Build the statement: DUMMY=DUMMY
- Choose [Simplex] estimation
- Click OK

The variable DUMMY is of no use to us except to trick **Nonlin** into accepting a null model statement. All the parameters (U and V) appear in the loss statement. Since the word ESTIMATE (estimated value) is not used in the loss statement, the model statement has no effect whatsoever on the minimization. Thus the dummy file should contain only one case to save time.

This particular function is from Rosenbrock (1960). We are using Simplex estimation to save space and because it generally does better with algebraic expressions that incur round-off error.

© 1989, SYSTAT, Inc.


```
Iteration       Loss          Parameter values
    0       .1620000D+01   .10000+00 .1000D+00
    1       .4509558D+00   .2406D+00 .5312D+00
    2       .1542587D+00   .5147D+00 .7342D+00
    3       .3195842D-01   .8708D+00 .9392D+00
    4       .8791500D-03   .1019D+01 .1010D+01
    5       .1285010D-05   .1001D+01 .1000D+01
    6       .1484245D-07   .1000D+01 .1000D+01
    7       .1244008D-10   .1000D+01 .1000D+01
    8       .6187833D-13   .1000D+01 .1000D+01

Dependent variable is    DUMMY
Final value of loss function is        0.000


Parameter        Estimate
        U           1.000
        V           1.000
```

7.13 A random model

You can use any operator or special variable in SYSTAT BASIC, including EOF, EOG, BOF, BOG, and CASE. What happens if you do this with the MEDICAL data?

- Select **Nonlin/Model...** from the **Stats** menu
- Build the statement: CARDIO=BETA*URN
- Click OK

Have fun.

7.14 Contouring the loss function

You can use the contouring routines in SYGRAPH to contour the parameters. Do not try to change the grid resolution. **Nonlin** generates a grid that looks best with the default [Cuts] setting. The grid is set to plot with even scale values at 10 tick marks, so do something like the following to produce a plot:

- Select **Plot/Contour** from the **Graph** menu
- Select the variables: LOSS *THETA2 *THETA1
- Set [X tick], [Y tick], and [Z tick] to 10 in the [Axes] dialog box

You might need to set [Z min] and [Z max] to get LOSS values in the range you want. **Nonlin** often saves a larger range of values than you need.

© 1989, SYSTAT, Inc.

For nonlinear least squares, you can set [Z tick] to 1, [Z min] to 0, and [Z max] to W, where

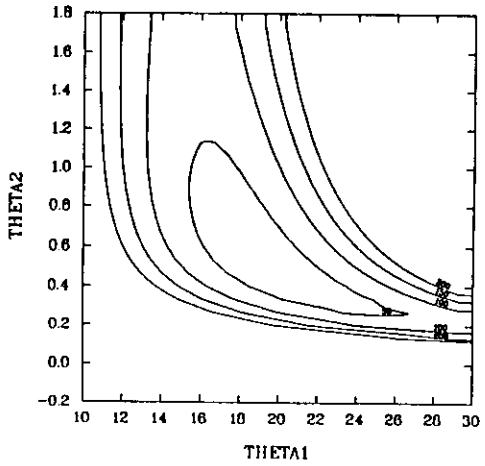
$$W = S\frac{1 + P}{(N - P)F}$$

Here, S is the loss function at minimum (sum of squares residual), F is the nominal F statistic at P and N–P degrees of freedom, P is 2 (2 parameters), and N is the number of cases.

Thus, you would do the following:

- Select 3-D Plot/3-D Plot from the Graph menu
- Select the variables: LOSS *THETA2 *THETA1
- Set [X tick] and [Y tick] to 10, [Z tick] to 1, [Z min] to 0, and [Z max] to 10000 in the [Axes] dialog box

Here is an example of several contours drawn from a data set in Bates and Watts (1988).



© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

8 Npar

Introduction 354
Rank (ordinal) data 354
Categorical (nominal) data 354
Robustness 355
An example 355
Usage 357
Friedman... 357
Kolmogorov-Smirnov... 358
One-sample tests 359
Two-sample tests 359
Lilliefors test 359
Kruskal-Wallis... 360
Sign test... 360
Wald-Wolfowitz... 361
Wilcoxon... 361
Computation 362
Missing data and ties 362
Algorithms 362
Examples 363
8.1 Sign test 363
8.2 Wilcoxon signed ranks test 364
8.3 Friedman two-way analysis of variance 365
8.4 Kruskal-Wallis one-way analysis of variance 365
8.5 Wald-Wolfowitz runs test 367
8.6 Kolmogorov-Smirnov one-sample test 368
8.7 Lilliefors test 370
8.8 Kolmogorov-Smirnov two-sample test 371

© 1989, SYSTAT, Inc.

Npar 8

Overview

Npar computes nonparametric statistics for groups of cases and pairs of variables. The tests offered by **Npar** fall under several categories. The following table provides a guide to their organization.

Data layout	*Test*
One sample	Kolmogorov-Smirnov
	Lilliefors
	Wald-Wolfowitz runs
Two independent samples	Kolmogorov-Smirnov
	Mann-Whitney U
Several independent samples	Kruskal-Wallis
Two related samples	Sign test
	Wilcoxon
Several related samples	Friedman

Many nonparametric statistics are computed elsewhere in SYSTAT. For example, **Corr** calculates matrices of coefficients like Spearman's rho, Kendall's tau-b, Guttman's mu2, and Goodman-Kruskal gamma. The **Graph** menu offers stem-and-leaf displays and box plots, which show medians, quartiles, and confidence intervals on medians. **Series** can do nonmetric smoothing. **Tables** can be used for chi-square tests of independence. **MDS** and **Cluster** work with nonmetric data matrices. Finally, the **Data** menu itself can rank and sort, and you can compute a variety of rank order statistics with DATA.

For references on the statistics included in this chapter, you might see Conover (1980), Hollander and Wolfe (1973), Mosteller and Rourke (1973), or Lehmann, (1975). Social scientists can consult the classic source of Siegel (1956), or a more recent text by Marascuilo and McSweeney (1977).

© 1989, SYSTAT, Inc.

Nonparametric statistics is a misnomer. The term is ordinarily used to describe a heterogeneous group of procedures that require relatively minimal assumptions about the shape of distributions underlying an analysis. Frequently, however, nonparametric models include parameters. These parameters are not necessarily ones like μ and σ, which we see in typical parametric tests based on normal theory, but they are parameters in a class of mathematical functions nonetheless.

In this context, a better term for nonparametric is *distribution-free*. That is, the data for this class of statistical tests are not assumed to follow a specific probability distribution. This does not mean, however, that we make *no* assumptions about distributions in nonparametric methods. For example, in the Mann-Whitney and Kruskal-Wallis test, we assume that the underlying populations are continuous and have the same shape.

Rank (ordinal) data

An aspect of many nonparametric tests is that they are invariant under *rank order* transformations of the data values. In other words, we may change actual data values as long as we preserve relative ranks, and the results of our hypothesis tests will not change. Data that can be replaced by rank order values without losing information are often called *rank* or *ordinal* data. For example, if we believe the list (–25, 54, 107.6, 3400) contains only ordinal information, then we can replace it with the list (1, 2, 3, 4) without loss of information.

Categorical (nominal) data

Some nonparametric methods are invariant under *permutation* transformations. That is, you can interchange data values and get the same results, provided you keep all cases with one value before transformation single valued after transformation. Data that you can treat like this are often called *categorical* or *nominal*. For example, if we believe the list (1, 1, 5, 5, 10, 10, 10) contains only nominal information, then we can replace it with the list (red, red, green, green, blue, blue, blue) without loss of information.

© 1989, SYSTAT, Inc.

Robustness

Sometimes, we may think our data contain more than nominal or ordinal information, but we wish to be extremely conservative. For example, our data may contain extreme outliers. We could eliminate these outliers, downweight them, or apply some nonlinear transformation to reduce their influence. An alternative, however, would be to use a nonparametric test based on ranks. If we can afford to lose some power by using a nonparametric test, we can gain robustness. If we find significant results with a nonparametric test, no skeptic can challenge us on the basis of scale artifacts or outliers. This is not to say that you should retreat to nonparametric methods every time you find a histogram that does not look normal. If you can find a simple normalizing transformation that works, such as logging the data, you will almost always be better off using normal parametric methods.

An example

The following table shows 20th century U.S. Presidents, their year of election, and their party. We are going to focus on the party variable. In the above terminology, party is nominal or categorical, because its values are simply names or categories for the specific cases.

YEAR	PRESDENT$	PARTY$
1900	McKinley	Republican
1904	Roosevelt	Republican
1908	Taft	Republican
1912	Wilson	Democrat
1916	Wilson	Democrat
1920	Harding	Republican
1924	Coolidge	Republican
1928	Hoover	Republican
1932	Roosevelt	Democrat
1936	Roosevelt	Democrat
1940	Roosevelt	Democrat
1944	Roosevelt	Democrat
1948	Truman	Democrat
1952	Eisenhower	Republican
1956	Eisenhower	Republican
1960	Kennedy	Democrat
1964	Johnson	Democrat
1968	Nixon	Republican
1972	Nixon	Republican
1976	Carter	Democrat
1980	Reagan	Republican
1984	Reagan	Republican
1988	Bush	Republican

© 1989, SYSTAT, Inc.

Hegel, Toynbee, and other philosophers and historians have claimed that political and intellectual history follows cycles. From their point of view, we might assume that the party of elected U.S. presidents would fluctuate somewhat regularly between Republican and Democrat.

An alternative view might be that fluctuations between parties of elected presidents are random, like the tossing of a coin. This null hypothesis need not imply that historical trends are random, but rather that the process represented by change of presidential party is random.

Let's test whether the runs of terms within the same party during this century are random. There are 23 terms in the series, with 9 runs of the same party. The Wald-Wolfowitz runs test shows that the chance of getting 9 or fewer runs of the given length by tossing a coin 23 times is .15. We cannot reject the null hypothesis that the runs pattern is random.

This doesn't mean, however, that Hegel and Toynbee are wrong. It is still possible that they might predict cycles of 4 years, or single terms, or even 8 year regular cycles. The runs test is not sensitive to the pattern of runs, only their number.

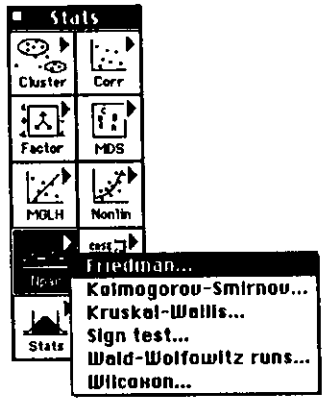
After Roosevelt's four consecutive terms, legislation was passed to limit the presidential office to a maximum of two terms. Did this action significantly reduce the length of runs in presidential party? The runs test indicates 4 runs in the 12 terms before and during Roosevelt's years and 6 runs in the following 11 terms. The probabilities associated with these two runs and their lengths are .07 and .95. It looks as if there is a tendency in the expected direction, but it is not significant.

To produce these results from the runs test, we made a variable with a value 1 for Republican and 2 for Democrat. It would not have mattered if we gave it a value of 30 for Republican and 500 for Democrat. The probabilities we found would have been the same.

© 1989, SYSTAT, Inc.

Usage

Nonparametric tests are computed by the Npar items of the same name, except Lilliefors' test, which is really a variant of the Kolmogorov-Smirnov test. The Mann-Whitney U test is computed by **Kruskal-Wallis...** automatically when the variables you select require it.



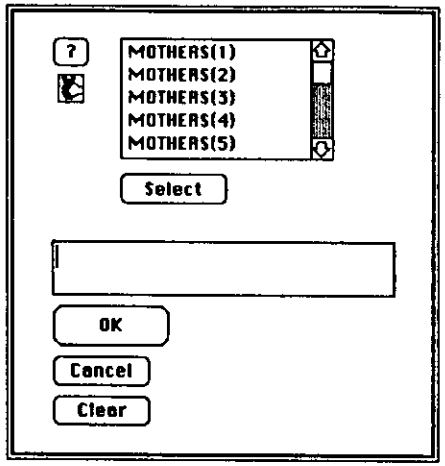
Press and hold the mouse button on **Npar**. A submenu pops up offering the various tests. Select the item you want and release the mouse button. The items offer similar dialog boxes; we show how to use each of them below.

Friedman...

The **Npar/Friedman...** dialog box is for related (dependent) samples. It offers procedures that are analogous to the paired *t*-test (Wilcoxon and Sign) and one-way repeated measures (Friedman) parametric tests.

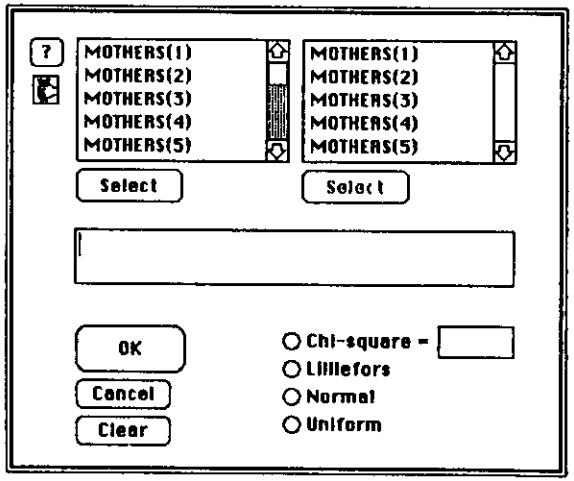
- Select **Npar/Friedman...** from the **Stats** menu
- Select two or more variables; all numeric variables are used by default
- Click OK

© 1989, SYSTAT, Inc.



Kolmogorov-Smirnov...

The **Npar/Kolmogorov-Smirnov...** dialog box can be used to compute one and two-sample Kolmogorov-Smirnov tests and Lilliefors tests. These are tests of the shape of discrete and continuous distributions. We give instructions for each type of test separately.



© 1989, SYSTAT, Inc.

One-sample tests

You can select one or more variables for one-sample tests. If you select more than one variable, separate tests are performed for each.

- Select **Npar/Kolmogorov-Smirnov...** from the **Stats** menu
- Select one or more variables from the left selection list
- Choose a distribution: Chi-square, Normal, or Uniform
  - Specify the degrees of freedom for the chi-square distribution
- Click OK

Two-sample tests

The data arrangement for two-sample tests can vary. If the sample sizes are the same, you can store each sample as a separate variable and select both variables from the left selection list. Both samples must be the same size after deleting missing values—that is, both variables must have the same number of *nonmissing* cases.

If the sample sizes differ, you must store all the data in a single variable and use a grouping variable to show which sample contains each case. Then, select the single data variable from the left selection list and the grouping variable from the right selection list.

- Select **Npar/Kolmogorov-Smirnov...** from the **Stats** menu
- Select two variables from the left list or one variable from the left list and a grouping variable from the right (see below)
- Choose a distribution: Chi-square, Normal, or Uniform
  - Specify the degrees of freedom for the chi-square distribution
- Click OK

If you do not select a grouping variable, all numeric variables are used by default. You must select a variable from the left list before you are allowed to select a grouping variable.

Lilliefors test

You can test normality without assuming a particular mean or standard deviation for the distribution by performing the Lilliefors test (Lilliefors, 1967). [Lilliefors] automatically standardizes the variables and tests whether the standardized versions are normally distributed. The test is not affected by the variable's original means and standard deviations.

- Select **Npar/Kolmogorov-Smirnov...** from the **Stats** menu
- Select one or more variables from the left selection list
- Choose [Lilliefors]
- Click OK

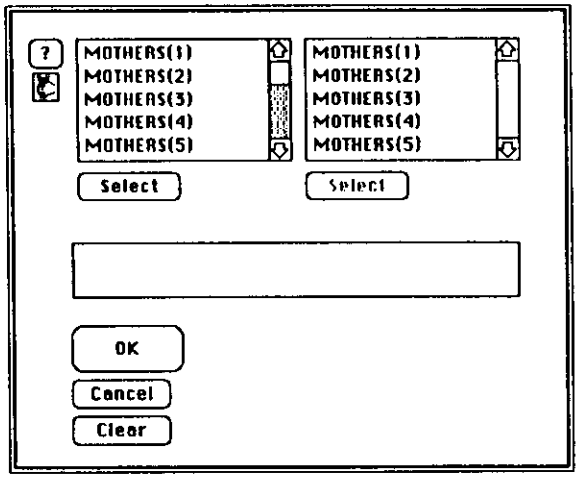
© 1989, SYSTAT, Inc.

If you select more than one variable, separate Lilliefors tests are performed for each.

Kruskal-Wallis...

The **Npar/Kruskal-Wallis...** dialog box is for tests on independent groups. It offers procedures that are analogous to the independent groups *t*-test (Mann-Whitney) and one-way ANOVA (Kruskal-Wallis). Character grouping variables are allowed.

- Select **Npar/Kruskal-Wallis...** from the **Stats** menu
- Select the variable(s) to be analyzed; all numeric variables are used by default
- Select a grouping variable
- Click OK



If you select more than one variable, separate ANOVAs are printed for each variable.

Sign test...

The dialog box for **Sign test...** is the same as that for **Friedman...**

- Select **Npar/Sign test...** from the **Stats** menu
- Select the variable(s) to be analyzed; all numeric variables are used by default
- Click OK

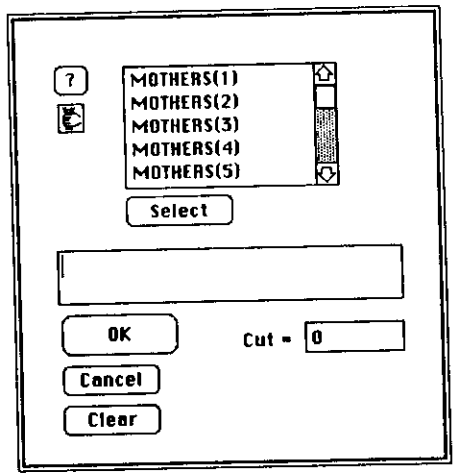
© 1989, SYSTAT, Inc.

If you select more than two variables, separate tests are performed for each pair of variables.

Wald-Wolfowitz...

The **Npar/Wald-Wolfowitz...** dialog box is for a runs test.

- Select **Npar/Wald-Wolfowitz...** from the **Stats** menu
- Select the variable(s) to be analyzed; all numeric variables are used by default
- Specify a cutpoint in the [Cut] box
- Click OK



If you select more than one variable, separate runs tests are performed for each.

Wilcoxon...

The dialog box for **Npar/Wilcoxon...** is the same as that for **Npar/Friedman...**

- Select **Npar/Wilcoxon...** from the **Stats** menu
- Select the variable(s) to be analyzed; all numeric variables are used by default
- Click OK

© 1989, SYSTAT, Inc.

Computation

Missing data and ties

The procedures in **Npar** delete missing data before computing statistics. In any procedures that involve pairs of variables (sign, Wilcoxon, Kolmogorov-Smirnov two-sample with separate variables), only pairs with nonmissing data for both items in the pair are used (pairwise deletion). Differences of zero and ties are handled by each procedure as follows:

Sign test...	Zero differences are deleted
Wilcoxon...	Zero differences are deleted, tied ranks averaged
Friedman...	Tied ranks are averaged
Kruskal-Wallis...	Tied ranks are averaged, and the test statistic is adjusted for the number of ties by the standard formula (e.g. Dixon,1981)

Algorithms

Probabilities for the Kolmogorov-Smirnov statistic for N<25 are computed with an algorithm in Pomeranz (1973). For larger samples, probabilities are calculated with an asymptotic negative exponential approximation.

Lilliefors probabilities are computed by a nonlinear approximation to Lilliefors' table. Dallal and Wilkinson (1986) recomputed Lilliefors' study using up to a million replications for estimating critical values. They found a number of Lilliefors' values to be incorrect. Consequently, the SYSTAT approximation uses the corrected values. The approximation discussed in Dallal and Wilkinson and used in SYSTAT differs from the tabled values by less than .01 and by less than .001 for $p<.05$.

© 1989, SYSTAT, Inc.

Examples

We use the following data from Siegel (1956) to illustrate several of the nonparametric tests. These are rankings of 20 mothers by 13 judges. The data are printed as ranks within rows, but for the procedures in **Npar**, this is not necessary. Each procedure ranks data before computing tests. This data set follows the revisions in Dixon (1981), with the exception of the 15th rank for the ninth judge, which has been adjusted for a tie.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	1	16	8	9	2	6	10	4	3	11	13	7	12	17	18	19	16	14	20
3	2	7	5	14	9	15	16	6	11	8	10	1	4	19	12	20	13	17	18
8	3	10	11	4	2	5	13	9	1	14	7	6	15	16	12	19	17	18	20
2	1	16	8	15	4	6	9	7	10	11.5	5	3	17	11.5	14	19	18	13	20
16	17	5	13	15	11	7	4	9	2	18	3	6	1	19	12	10	8	14	20
12	9	14	6	7	2	3	10	5	4	17	8	1	15	13	16	18	11	20	19
11	2	13	10	7	3	4	14	6	5	17	9	1	12	8	16	20	15	18	19
9.5	2	15.5	6	5	7	8	11	9.5	3	13	4	1	14	12	15.5	20	19	17	18
2	4	16	3	10	6	14	17	15	7	19	9	1	8	5	13	11	18	12	20
11	14	12	8	7	2	5	10	3	4	13	9	1	18	6	15	19	16	17	20
5	3	13	2	8	1	9	12	4	6	14	10	11	7	15	18	16	17	19	20
8	1	13	3	5	2	14	9	6	10	15	11	19	4	7	12	18	17	16	20

8.1 Sign test

Npar/Sign test... computes a sign test for selected pairs or all possible pairs of numeric variables in a SYSTAT file. For each pair, the difference between values on each case is calculated and the number of positive and negative differences is printed. The lesser of the two types of differences (positive or negative) is then compared to the total number of nonzero differences. From this comparison, the probability is computed according to the binomial (for a total less than or equal to 25) or a normal approximation to the binomial (for a total greater than 25). A correction for continuity (.5) is added to the normal approximation's numerator and the denominator is computed from the null value of .5. The large sample test is thus equivalent to a chi-square test for an underlying proportion of .5.

- Select **Npar/Sign test...** from the **Stats** menu
- Select the variables to be analyzed: MOTHERS(1), MOTHERS(2)
- Click OK

© 1989, SYSTAT, Inc.

```
Sign Test Results

Counts of Differences (Row Variable Greater Than Column)

                MOTHERS(1)  MOTHERS(2)

MOTHERS(1)              0           9
MOTHERS(2)              4           0

Two-Sided Probabilities for Each Pair of Variables

                MOTHERS(1)  MOTHERS(2)

MOTHERS(1)          1.000
MOTHERS(2)           .267       1.000
```

8.2 Wilcoxon signed ranks test

If you have data that are not binary, the sign test can be a bit drastic. A comparable test that takes into account the size of rank order differences within pairs is the Wilcoxon signed ranks test. In this test, the absolute values of all nonzero differences are ranked. For ties, the average rank is assigned. The sum of the ranks associated with positive differences and the sum of ranks associated with negative differences is then computed. The test statistic is the lesser of these two sums. Two-tail probabilities are computed from an approximate normal variate constructed from this statistic (e.g. Marascuilo and McSweeney, 1977, p. 338). Note that the Wilcoxon test computes the sum of the ranked differences rather than merely the number of positive and negative differences as in the sign test.

- Select **Npar/Wilcoxon...** from the **Stats** menu
- Select the variables to be analyzed: MOTHERS(1), MOTHERS(2)
- Click OK

```
Wilcoxon Signed Ranks Test Results

  Counts of Differences (Row Variable Greater Than Column)

                MOTHERS(1)  MOTHERS(2)

MOTHERS(1)              0           9
MOTHERS(2)              4           0

  Z = (Sum of Signed Ranks)/Square Root(Sum of Squared Ranks)

                MOTHERS(1)  MOTHERS(2)

MOTHERS(1)          1.000
MOTHERS(2)         -1.929       1.000

  Two-Sided Probabilities Using Normal Approximation

                MOTHERS(1)  MOTHERS(2)

MOTHERS(1)          1.000
MOTHERS(2)           .054       1.000
```


© 1989, SYSTAT, Inc.

8.3 Friedman two-way analysis of variance

The Friedman two-way analysis of variance is useful for analyzing rankings of a set of objects by multiple judges. The ratings of mothers from Siegel is a typical example. Other applications are rankings of wines by experts, artistic judgments, and consumer ratings. The Friedman statistic is used to test the hypothesis that there is no systematic variation in the rankings across objects (i.e., that the judges do not agree predictably in their ratings). Since we want to test all the (numeric) variables in the file (the default), we do not need to select any variables.

- Select **Npar/Friedman...** from the **Stats** menu
- Click OK

The Kendall coefficient of concordance is an estimate of the average correlation among the judges' ratings of the mothers.

```
Friedman Two-Way Analysis of Variance Results for  13 Cases

  Variable     Rank Sum

  MOTHERS(1)       93.5
  MOTHERS(2)       61.0
  MOTHERS(3)      153.0
  MOTHERS(4)       87.0
  MOTHERS(5)      111.0
  MOTHERS(6)       57.0
  MOTHERS(7)      103.0
  MOTHERS(8)      143.0
  MOTHERS(9)       92.5
  MOTHERS(10)      76.0
  MOTHERS(11)     181.5
  MOTHERS(12)     110.0
  MOTHERS(13)      71.0
  MOTHERS(14)     141.0
  MOTHERS(15)     163.5
  MOTHERS(16)     189.5
  MOTHERS(17)     226.0
  MOTHERS(18)     202.5
  MOTHERS(19)     214.0
  MOTHERS(20)     254.0

Friedman Test Statistic = 146.72
Kendall Coefficient of Concordance =    .594

Probability is  .000 Assuming Chi-Square Distribution with 19 DF
```

8.4 Kruskal-Wallis one-way analysis of variance

The Kruskal-Wallis one-way analysis of variance is an ANOVA of ranked data. The test substitutes an approximation to a chi-square for an F-statistic.

Before we do a Kruskal-Wallis test on the CRIME data, let's compute a regular one-way ANOVA comparing 1970 murder rates across 8 Census regions of the United States. The "cases" are states and the data are murders per 100,000 population.

© 1989, SYSTAT, Inc.

- Select **MGLH/Fully factorial (M)ANOVA...** from the **Stats** menu
- Select the [Dependent] variable: MURDER
- Click [Factors] and select the factor, specifying the number of levels: REGION (8 levels)
- Click OK

```
                    Analysis of Variance

Source    Sum-of-squares    DF  Mean-Square     F-Ratio       P

REGION         400.134      7       57.162       7.378     0.000
ERROR          325.407     42        7.748
```

From the size of the REGION sum of squares, it would seem that the variation in MURDER is substantial across regions. Residual plots for this analysis show problems with heteroscedasticity (unequal variances) in MURDER across regions. We should probably transform the data and reanalyze.

A drastic transformation is simply to rank the data. That is what the Kruskal-Wallis test does. It *is* drastic, however, so you should consider other linearizing transformations before giving up on regular ANOVA (see Box and Cox, 1964, for example, or Mosteller and Tukey, 1977).

- Select **Npar/Kruskal-Wallis...** from the **Stats** menu
- Select the variable to be analyzed: MURDER
- Select a grouping variable: REGION
- Click OK

```
Kruskal Wallis One-way Analysis of Variance for  50 cases
  Dependent variable is  MURDER
  Grouping variable is   REGION

    Group       Count    Rank Sum

        1.000       6        46.5
        2.000       3        73.5
        3.000       5       126.0
        4.000       7        94.5
        5.000       8       312.5
        6.000       8       310.5
        7.000       8       191.0
        8.000       5       120.5

Kruskal-Wallis test statistic =    27.42
Probability is 0.000 assuming chi-square distribution with  7 DF
```

Still seems significant. And with this analysis we don't have to worry about departures from normality.

© 1989, SYSTAT, Inc.

Both of these analyses, however, violate basic assumptions. First, they lack a plausible probability model. The 50 cases are an exhaustive sample—in fact, no sample at all. Second, the MURDER values have a spatial distribution. Adjacent states have similar values for all sorts of reasons. Therefore, the cases are not independent of one another.

Some researchers fall back on nonparametric tests as a panacea for "bad data." Keep in mind that there are explicit assumptions required for nonparametric analyses as for any other tests based on a probability model. Typically, the Kruskal-Wallis model requires that all possible partitions of the data (ranks) consistent with the observed sample sizes be equally likely in the absence of the effect being tested. If some of these partitions are counterfactual or inconsistent with common sense, then a probability model is inappropriate.

Finally, beware of using nonparametric procedures to rescue bad data. These procedures were in most cases designed to apply to data that were categorical or ranked in the first place, such as rank judgments and binary data. If you have data that violate distributional assumptions for linear models, for example, you should consider transformations or robust models before retreating to nonparametrics.

8.5 Wald-Wolfowitz runs test

The Wald-Wolfowitz runs test detects serial patterns in a run of numbers. For example, if you toss a coin 20 times, you might get the following sequence:

HTHHHHHHHHTTTTTTTTTT

A total of 9 heads and 11 tails in 20 tosses doesn't look very surprising for a fair coin, but the long runs of each kind look uncanny. The sign test assesses the first type of nonrandom behavior (total number of heads versus tails). The runs test measures the second type of nonrandom behavior for a dichotomous, or binary, variable. In this example, we find 4 runs of head or tail (even a single head or tail counts as a run).

Sometimes you have a continuous variable you would like to examine for runs above and below some cutpoint. For example, you may want to see if residuals from a regression tend to fluctuate in patterns above and below the zero line. To do this, use the [Cut] option of the runs test. If you do not specify a cutpoint, it is assumed to be 0.

© 1989, SYSTAT, Inc.

Here is an example for the SALES data. We will cut MRCHDSE at its median (you can use **Graph/Stem** to find the median of a variable) and see if states with higher merchandise sales are bunched together in the file.

- Select **Npar/Wald-Wolfowitz...** from the **Stats** menu
- Select the variable to be analyzed: MRCHDSE
- Set [Cut] to 1743
- Click OK

```
Wald-Wolfowitz Runs Test Using Cutpoint-   0.174E+04

  Variable  Cases LE Cut  Cases GT Cut   Runs        Z     Probability(2-Tail)

MRCHDSE          24            23        19      -1.620          .105
```

There does not appear to be a significant pattern of runs above and below the median. Be careful when you use a cutpoint on a continuous variable, however. Your conclusions can change depending on the cutpoint you use. Here is what happens if we use the mean.

- Select **Npar/Wald-Wolfowitz...** from the **Stats** menu
- Select the variable to be analyzed: MRCHDSE
- Set [Cut] to 2522
- Click OK

```
Wald-Wolfowitz Runs Test Using Cutpoint-   0.252E+04

 Variable Cases LE Cut Cases GT Cut  Rns     Z   Probability(2-Tail)

 MRCHDSE     33      14    15    -2.004     .045
```

It begins to look significant. Keep in mind some of the warnings above about using nonparametrics on continuous data. They are no panacea.

8.6 Kolmogorov-Smirnov one-sample test

The Kolmogorov-Smirnov test and its generalizations are among the handiest of distribution-free tests. It is particularly useful for testing the shape and location of a sample distribution. The test statistic is based on the maximum difference between two cumulative distribution functions. In the one-sample test, one of the CDFs is continuous and the other discrete. Thus, it is a companion test to a probability plot.

© 1989, SYSTAT, Inc.

As an example, let's use the SYSTAT random number generator to make a normally distributed random number and then test it for normality with the KS command. Also, to make a point, we use the variable Z as our normal random number and the variable ZS as a standardized copy of Z. This may seem strange, since normal random numbers are expected to have a mean of zero and a standard deviation of one. This is not exactly true in a sample, however, so we standardize the observed values to make a variable that has exactly a mean of zero and a standard deviation of one.

- Clear the Editor by selecting **New** from the **File** menu
- Name the first variable Z

- Select **Fill worksheet...** from the **Editor** menu
- Specify 50 cases to be filled
- Click OK

- Select **Math...** from the **Editor** menu
- Fill a variable Z with normal random data: Set Z to ZRN
- Click OK

- Select **Math...** from the **Editor** menu
- Make a copy of Z: Set Z to ZS
- Click OK

- Select **Standardize...** from the **Data** menu
- Select the variable to be standardized: ZS
- Click OK

The variable ZS is standardized in the new file you specify. Now, use **File/Open...** to open the file you just saved with **Data/Standardize...** This new file has the variable of normal random data, Z, and a standardized version of the same variable, ZS.

Next, let's use **Stats** to see the mean and standard deviation of our two variables. Remember, if you correlated these two variables with **Corr**, the Pearson correlation would be 1. Only their mean and standard deviations differ.

- Select **Stats/Statistics...** from the **Stats** menu
- Select the variables to be analyzed: Z and ZS
- Click OK

© 1989, SYSTAT, Inc.

```
Total Observations:   50

                          Z          ZS

N of Cases               50          50
Minimum              -2.491      -1.998
Maximum               4.249       3.585
Mean                 -0.078       0.000
Standard Dev          1.207       1.000
```

Now, we test Z for normality with **Npar**. Since Z and ZS are the only variables in the file and we want to test both of them, we need not select variables.

- Select **Npar/Kolmogorov-Smirnov...** from the **Stats** menu
- Choose a distribution: [Normal]
- Click OK

```
Kolmogorov-Smirnov One Sample Test Using Standard Normal Distribution

  Variable     N of Cases   Maxdif   Probability (2-TAIL)

         Z             50     .118                .490
        ZS             50     .110                .582
```

Why are the probabilities different? The Kolmogorov-Smirnov one-sample test pays attention to the shape, location, and scale of the sample distribution. Z and ZS have the same shape in the population (because they are both normal). Since ZS has been standardized, however, it has a different location (because we changed the mean) and scale (because we changed the standard deviation).

Thus, you should never use the Kolmogorov-Smirnov test with the normal distribution on a variable you have standardized. The probability printed for ZS (.582) is misleading. To repeat, if you use the [Normal], [Uniform], or [Chi-square] options with the Kolmogorov-Smirnov test, you are assuming that the variable you are testing has been randomly sampled from a standard normal, uniform (0 to 1), or chi-square (with stated degrees of freedom) population.

8.7 Lilliefors test

Here we perform a Lilliefors test on the same variables. Note that the [Lilliefors] option automatically standardizes the variables you select and tests whether the standardized versions are normally distributed.

- Select **Npar/Kolmogorov-Smirnov...** from the **Stats** menu
- Click the [Lilliefors] option
- Select the variables to be analyzed: Z, ZS

© 1989, SYSTAT, Inc.


```
Kolmogorov-Smirnov One Sample Test Using Standard Normal Distribution

  Variable     N of Cases   Maxdif   Lilliefors Probability (2-Tail)

         Z             50     .110                .131
        ZS             50     .110                .131
```

Notice the probabilities are smaller this time even though MAXDIF is the same as before. The probability values for Z and ZS (.131) are the same because this test pays attention only to the shape of the distribution and not to the location or scale. Neither significantly differs from normal. This example was constructed to contrast the [Normal] and [Lilliefors] options. Many statistical package users do a Kolmogorov-Smirnov test for normality on their standardized data without realizing that they should instead do a Lilliefors test. For more information, see the "Algorithms" section under "Computation."

One last point. The [Lilliefors] option is handy for residual analysis in regression. Just standardize your residuals and use **Npar** to test them for normality. If you do this, you should always look at the corresponding normal probability plot.

8.8 Kolmogorov-Smirnov two-sample test

The Kolmogorov-Smirnov two-sample test measures the discrepancy between two sample cumulative distribution functions. It is a companion test to a Quantile-Quantile plot.

We are going to test whether Z and ZS have the same mean, standard deviation, and shape. For normal distributions, location and scale totally determine shape. For other distributions, this may not be true. When you use the KS two-sample test, you assume both samples came from exactly the same distribution. It is possible, for example, that two samples correlate perfectly and yet significantly differ on a KS test because the Pearson correlation equates location and scale.

Note that we do not need to select variables; the default (all numeric variables in the file) is fine.

- Select **Npar/Kolmogorov-Smirnov...** from the **Stats** menu
- Click OK

© 1989, SYSTAT, Inc.

```
Kolmogorov-Smirnov Two Sample Test Results

 Maximum Differences for Pairs of Variables

                      Z          ZS

       Z             .000
      ZS             .120        .000

 Two-Sided Probabilities

                      Z          ZS

       Z            1.000
      ZS             .823       1.000
```

The test does not significantly distinguish Z and ZS.

Different sample sizes

If you want to compare two samples with different numbers of cases, you need to use a grouping variable.

- Select **Npar/Kolmogorov-Smirnov...** from the **Stats** menu
- Select the variable to be analyzed: Z
- Select a grouping variable
- Click OK

In this case, Z is a single variable with values for each sample represented by the grouping variable. You may select only one variable and one grouping variable. The output is labeled with the values of the groups contained in the grouping variable.

© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

9 Series

Introduction 376
Smoothing 376
Moving averages 377
Weighted running smoothing 380
Running median smoothers 380
LOWESS 381
And the winner is... 382
ARIMA modeling and forecasting 383
Autocorrelation 385
Autoregressive models 386
Moving average models 388
ARMA models 388
Identifying models 389
Estimating the ARIMA model 392
Problems with forecast models 392
Seasonal decomposition and adjustment 393
Exponential smoothing 394
Fourier analysis 395

Usage 397
Transform... 398
ID... 401
Clear 401
Case plot... 402
ACF plot... 403
PACF plot... 403
CCF plot... 404
ARIMA... 404
Smooth... 407
LOWESS... 407
Exponential... 408
Seasonal... 410
Fourier... 412

Computation 414
Missing data 414
Algorithms 414

Examples 416
9.1 Series plots and differencing 416
9.2 ACF, PACF, CCF plots 419
9.3 ARIMA model identification 425
9.4 ARIMA with backcasting 430
9.5 ARIMA with forecasting 433
9.6 Exponential smoothing 434
9.7 Seasonal decomposition and adjustment 437
9.8 Fourier analysis 439

© 1989, SYSTAT, Inc.

Series 9

Overview

Series implements a wide variety of time series models, including linear and nonlinear filtering, Fourier analysis, seasonal decomposition, nonseasonal and seasonal exponential smoothing, and the Box-Jenkins approach to nonseasonal and seasonal ARIMA. In addition, interactive transformation and graphics capabilities can help you choose appropriate models without wading through lengthy output.

© 1989, SYSTAT, Inc.

Introduction

Time series analysis can range from the purely exploratory to the confirmatory testing of formal models. Series encompasses both exploratory and confirmatory methods. Among the exploratory methods are smoothing and plotting. Among confirmatory models are two general approaches: time domain and frequency domain. In time domain models, we examine the behavior of variables over time directly. In frequency domain models, we examine frequency (periodic) components contributing to a time series.

Time domain (autoregressive, moving average, and trend) models represent a series as a function of previous points in the same series or as a systematic trend over time. Time domain models can fit complex patterns of time series with just a few parameters. Makridakis, Wheelwright and McGee (1983), McCleary and Hay (1980), and Nelson (1973) introduce these models, while Box and Jenkins (1976) provide the primary reference for ARIMA models.

Frequency domain (spectral) models decompose a series into a sum of sinusoidal (waveform) elements. These models are particularly useful when a series arises from a relatively small set of cyclical functions. Bloomfield (1976) introduces these models.

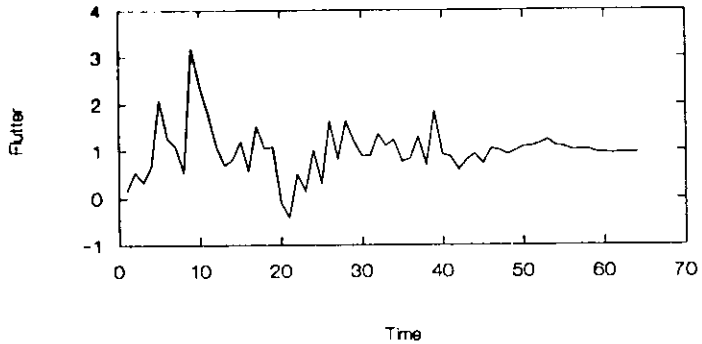
In this introduction, we will discuss exploratory methods (smoothing), time domain models (ARIMA, seasonal decomposition, exponential smoothing), and frequency domain (Fourier) models.

Smoothing

Sometimes, with a "noisy" time series, we seek some sort of smooth trend but have no idea what type of function generated the series. The incidence of daily runaways from a low security boys' reformatory, daily fluctuations in body temperature of cows, the increasingly erratic eruptions of the "Old Faithful" Geyser, all could give us valuable information if we could detect patterns. Smoothing is a complex topic whose applications exceed space here, so you should consult Velleman and Hoaglin (1981) or Bloomfield (1976) for more complete discussions.

© 1989, SYSTAT, Inc.

The following plot gives a typical "free decay record" from a "flutter test" of an aircraft wing. (Bennett & Desmarais, 1975). It shows amplitude of vibration versus time in an aircraft component. Although the model for these data is known, we are going to try to recover a smooth series without using this information. The data are in the file AIRCRAFT.



Moving averages

A variety of techniques can smooth, or filter the noise from, this series. One of the simplest is a moving average. If a data point consists of a smooth component plus random error, then if we average several points surrounding a point, the errors should tend to cancel each other out.

Here are two possible moving averages three and four points wide. The window shows which points are being averaged. The boldface shows which point in the series is replaced with the average.

© 1989, SYSTAT, Inc.

Three point window

Series	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9
Window	y_1	y_2	y_3						
		y_2	y_3	y_4					
			y_3	y_4	y_5				
				y_4	y_5	y_6			
					y_5	y_6	y_7		
						y_6	y_7	y_8	
							y_7	y_8	y_9
New series	y_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	y_9

Four point window

Series	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
Window	y_1	y_2	y_3	y_4						
		y_2	y_3	y_4	y_5					
			y_3	y_4	y_5	y_6				
				y_4	y_5	y_6	y_7			
					y_5	y_6	y_7	y_8		
						y_6	y_7	y_8	y_9	
							y_7	y_8	y_9	y_{10}
New series	y_1	y_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	y_{10}

Notice that the four point window does not have a point in the series at its center. Consequently, we replace the right point of the two in the middle with the average of the four points. This rule is followed for all even windows except two point windows. Two point windows can thus be used to shift asymmetrical smoothings back to the left.

If you prefer algebra, then the following description shows how the three point window smooths y into x.

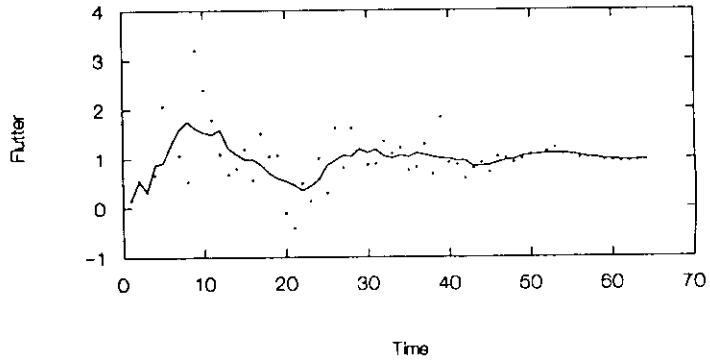
© 1989, SYSTAT, Inc.

$$x_1 = y_1$$

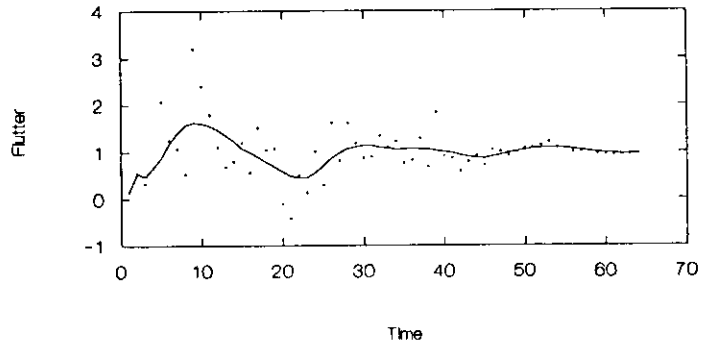
$$x_2 = \frac{y_1 + y_2 + y_3}{3}$$

$$x_3 = \frac{y_2 + y_3 + y_4}{3}$$

Notice also that the first and last points in the series are unchanged by the three point window of moving averages. The four point window leaves the first two and last two points unchanged. Let's try a seven point moving average on this series.



We now have a considerably smoother display. Let's smooth this smoothed series with a 4 point moving average. This should take out some of the shorter "jitters" at the end of the series.



© 1989, SYSTAT, Inc.

This is even smoother. We chose the lengths of the window by trial and error after looking at the data to see how much they "jitter" to the left and right of each point relative to the overall pattern of the series. You will do better if you know something about the function generating the data.

Weighted running smoothing

If you know something about filter design (see Bloomfield, 1976), you can construct a more general linear filter by using the [Weight] option. We have already done 7 and 4 point moving averages with equal weights.

Our earlier smoothings used even weights of 1 for each member in the window since we did not specify otherwise. We could, however, set these weights to any real number; e.g., 1,2,1. Some of you may recognize these as *Hanning* weights (Chambers, 1977; Velleman & Hoaglin, 1981). It is possible to show algebraically that weighting by (1,2,1) in a three observation window is the same as smoothing twice with equal weights in a two observation window. The [DWLS] smoothing method in SYGRAPH is a form of weighting in which weights are determined by distance weighted least squares.

Running median smoothers

Now, let's look at another smoother: running medians. Sometimes it's handy to have a more robust filter when you suspect the data do not contain Gaussian noise. You can choose this filter with the [Median] option of the running smoother. It works like the [Mean] option, except the values in the series are replaced by the median of the window instead of the mean.

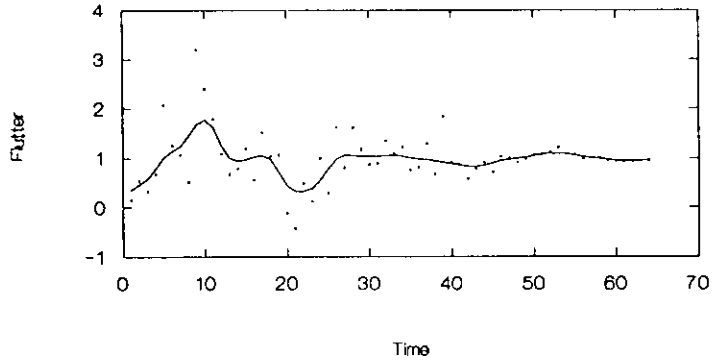
Can you see why running mean and running median smoothers with a window of two are the same?

We can use combinations of these smoothers to construct more complex nonlinear filters. The following sequence of smoothings comprises a nonlinear filter because it doesn't involve a simple weighted average of the values in a window (except for the final Hanning step). It uses a combination of running medians instead.

© 1989, SYSTAT, Inc.

Running median smoother, window 4
Running median smoother, window 2
Running median smoother, window 5
Running median smoother, window 3
Running means smoother, window 3, weights 1, 2, 1

You can read about this filter (called 4253H) in Velleman and Hoaglin (1981). It is due to the work of Tukey (1977). It happens to be a generally effective compound smoother because it clears outliers out of the sequence in the early stages and polishes up the smooth later. Velleman and Hoaglin use this smoother twice on the same data by smoothing the data, smoothing the residuals from this smooth, and adding the two together. You can do this by clicking [Save file] with the last smoothing to save the smoothed values into a SYSTAT file. You can then merge the files and compute residuals. In the final step, you can smooth the residuals. Here is the resulting plot of the wing flutter data after the five smooths above.



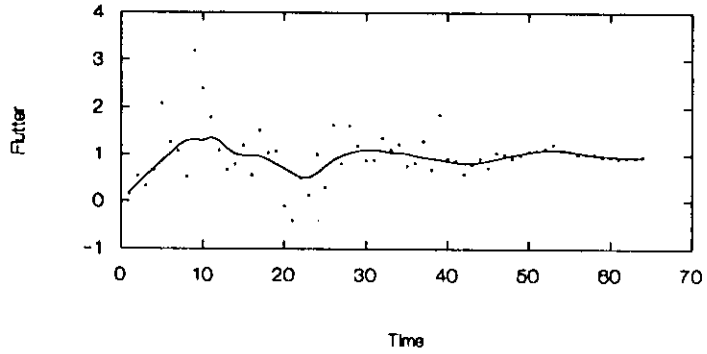
LOWESS

Cleveland (1979) presented a method for smoothing values of Y paired with a set of ordered X values. Chambers *et al.* (1983) introduce this technique and present some clear examples. If you are not a statistician, by the way, and want a glimpse of some of the most exciting things professional statisticians are doing today, read the Chambers *et al.* book (and Velleman & Hoaglin if you don't know about Tukey's work). Both are available in paperback.

© 1989, SYSTAT, Inc.

Scatterplot smoothing allows you to look for a functional relation between Y and X without prejudging its shape (or its monotonicity). The method for finding smoothed values involves a locally weighted robust regression. SYSTAT implements Cleveland's LOWESS algorithm.

Series implements LOWESS on equally spaced data values. You can also use LOWESS on scatterplots of unequally spaced data values in **Graph**. Here is our flutter data smoothed by LOWESS. We used a *tension* value of .18 to get more of the local detail.



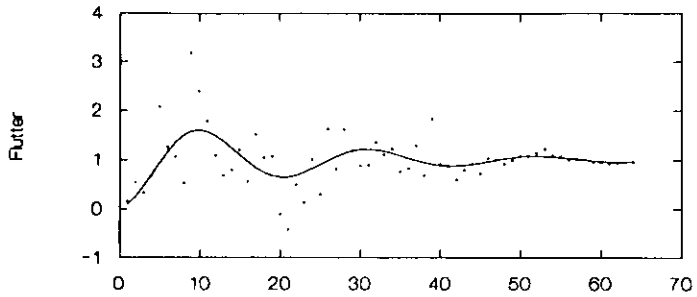
And the winner is...

Here is the actual function used to generate the data:

$$Y(t) = 1 - e^{-.05t}\cos(.3t)$$

where t =1, 2, ..., 64 (the index number of the series). We added normal (Gaussian) noise to this function in inverse proportion to the square root of t. We leave it to you engineers to design an optimal filter for the [Weight] option after looking at the noise distribution in the plot. Here is the generating function on the data:

© 1989, SYSTAT, Inc.



Is there a winner? For Gaussian data like these, it is hard to beat running means. Running medians and LOWESS do extremely well on non-Gaussian data, however, because they are less susceptible to outliers in the series. You will also find that exploratory smoothing requires a lot of fine tuning with window widths (*tension*) and with weights.

ARIMA modeling and forecasting

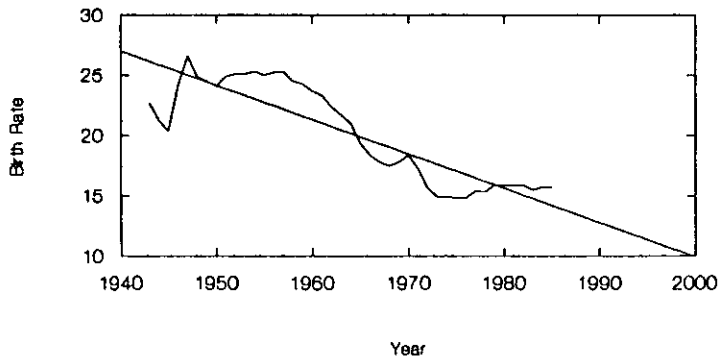
The following data show the U.S. birth rate (per 1000) for several decades during and following World War II. They were compiled from federal statistics, principally the U.S. census.

YEAR	RATE
1943	22.7
1944	21.2
1945	20.4
1946	24.1
1947	26.6
1948	24.9
1949	24.5
1950	24.1
1951	24.9
1952	25.1
1953	25.1
1954	25.3
1955	25.0
1956	25.2
1957	25.3
1958	24.5
1959	24.3
1960	23.7
1961	23.3
1962	22.4
1963	21.7

© 1989, SYSTAT, Inc.

1964	21.0
1965	19.4
1966	18.4
1967	17.8
1968	17.5
1969	17.8
1970	18.4
1971	17.2
1972	15.6
1973	14.9
1974	14.9
1975	14.8
1976	14.8
1977	15.4
1978	15.3
1979	15.9
1980	15.9
1981	15.9
1982	15.9
1983	15.5
1984	15.7
1985	15.7

These data are a *time series*, because they comprise values on a variable distributed across time. How can we use these data to forecast birth rates up to the year 2000? A popular statistical method for such a forecast is linear regression. Let's try it. Here is a plot of birth rates against year with the least squares line. I have connected the data points so that you can see the series more clearly.



© 1989, SYSTAT, Inc.

What's wrong with this forecasting method? You may want to read the "Introduction" section in the **MGLH** chapter if you haven't already. There, we discussed assumptions needed for estimating a model using least squares. We can legitimately fit a line to these data by least squares for the explicit purpose of getting predicted values on the line as close as possible, on average, to observed values in the data. In forecasting, however, we want to use a fitted model to extrapolate beyond the series. The fitted linear model is:

$$\text{RATE} = 579.342 - .285 \times \text{YEAR}$$

If we want our estimates of the slope and intercept in this model to be unbiased, we need to assume that the errors (ε) in the population model are independent of each other and of YEAR. Does our data plot give us any indication of this?

On the contrary, it appears from the data that the randomness in this model is related to YEAR. Take any two adjacent years' data. On average, if there is an underprediction one year, there will be an underprediction the next. If there is overprediction one year, there is likely to be overprediction the next. These data clearly violate the assumption of independence in the errors.

Autocorrelation

There is a statistical index that reveals how correlated the residuals are. It is called the *autocorrelation*. The first-order autocorrelation is the ordinary Pearson correlation of a series of numbers with the same series shifted by one observation ($y_2, y_1; y_3, y_2; \ldots; y_n, y_{n-1}$). In our residuals from the linear model, this statistic is .944. If you remember about squaring correlation coefficients to reveal proportion of variance, this means that over 88 percent of the variation in error from predicting one year's birth rate can be accounted for by the error in predicting the previous year's.

© 1989, SYSTAT, Inc.

The second-order autocorrelation is produced by correlating the series $(y_3, y_1; y_4, y_2; \ldots; y_n, y_{n-2})$. Computing this statistic involves shifting the series down *two* years. As you may now infer, we can keep shifting and computing autocorrelations for as many years as there are in the series. There is a simple graphical way to display all these autocorrelations. It looks like a bar graph of the autocorrelations sequenced by year, or index in the series. The top bar is the first autocorrelation (.944). The next highest bar is the second, and so on. Here it is:

```
Plot of      RATE
Number of cases =   43
Mean of series =      20.186
Standard deviation of series =        4.059

Plot of Autocorrelations

 Lag   Corr     SE -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
                    |----|----|----|----|----|----|----|----|----|----|
   1    .953   .152                        (   |******)***************
   2    .896   .256                    (       |**********)**********
   3    .854   .321                (           |**************)******
   4    .806   .370               (            |****************)**
   5    .741   .409              (             |***************** )
   6    .671   .439             (              |***************    )
   7    .597   .462           (                |*************        )
   8    .515   .480           (                |***********          )
   9    .428   .492         (                  |*********             )
  10    .335   .501   (                        |********                )
  11    .234   .506   (                        |*****                   )
  12    .134   .509   (                        |***                     )
  13    .039   .509   (                        |                        )
  14   -.051   .509   (                      . |                        )
  15   -.130   .510   (                   ***  |                        )
```

This autocorrelation plot tells us about all the autocorrelations in the residuals from the linear model. As you can see, there is a strong dependence in the residuals. As we shift the series far enough back, the autocorrelations become negative, because the series crosses the prediction line and the residuals become negative. Over the entire series, there are three crossings and three corresponding shifts in sign among the autocorrelations.

Autoregressive models

We would have the same serial correlation problem if we refined our model to include a quadratic term:

$$\text{RATE} = \beta_0 + \beta_1 \text{YEAR} + \beta_2 \text{YEAR}^2 + \varepsilon$$

You can try this model with **MGLH**, but you will find a large autocorrelation in the residuals even though the curve fits the data more closely. How can we construct a model that includes the autocorrelation structure itself?

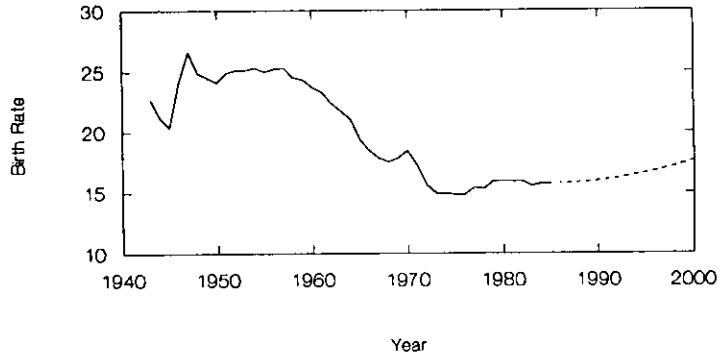
© 1989, SYSTAT, Inc.

The *autoregressive model* does this:

$$\text{RATE}_i = \beta_0 + \beta_1 \text{RATE}_{i-1} + \varepsilon_i$$

Notice that this model expresses a year's birth rate as a function of the previous year's birth rate—not as a function of YEAR. Time becomes a sequencing variable, not a predictor.

To fit this, we fit an AR(1) model with the ARIMA procedure. You can read more about it in the technical section of this chapter. Here is the result, with forecasts extending to the year 2000:



The forecasted values are represented by the dotted line. Unlike the regression model forecast, the autoregressive forecast begins at the last birth rate value and drifts back toward the mean of the series. This forecast behavior is typical of this particular model, which is often called a *random walk*.

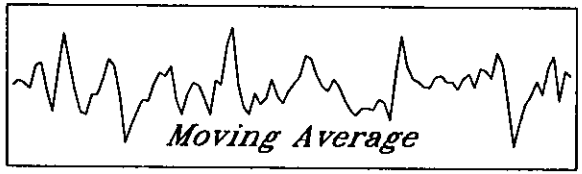
© 1989, SYSTAT, Inc.

Moving average models

There is another series model that can account for fluctuations across time. The *moving average model* looks like this:

$$y_i = \varepsilon_i - \beta_1 \varepsilon_{i-1}$$

This models a series as a cumulation of random shocks or disturbances. If this model represented someone's spending habits, for example, then whether the person went on a spending spree one day would depend on whether he or she went on one the day before. Unlike the autoregressive model, which represents an observation as a function of previous observations' *values*, the moving average model represents an observation as a function of the previous observations' *errors*. So you can see the difference between the two, here are examples of first-order autoregressive, or AR(1), and moving average, or MA(1), series:



ARMA models

Autoregressive and moving average models can be mixed to make autoregressive-moving average models. They can be mixed with different orders, e.g. AR(2) plus MA(1), which is often expressed as ARMA(2,1). You will have to refer to a text on forecasting to see instances of these more complicated models. You could visually add the two sample series above, however, to see how an ARMA(1,1) model would look.

© 1989, SYSTAT, Inc.

Identifying models

Before you can fit an AR, MA, or ARMA model, you need to identify which model is appropriate for your series. You can look at the series plot to find distinctive patterns, as in the figure contrasting AR(1) and MA(1) directly above. Real data seldom fit these ideal types as clearly, however. There are several powerful tools that distinguish these families of models. We have already seen one: the *autocorrelation function plot* (ACF). The *partial autocorrelation function plot* (PACF) provides additional information about serial correlation. To identify models, we use both of these plots.

Stationarity

Before doing these plots, however, you should be sure the series is *stationary*. This means:

1) **The mean of the series is constant across time.** You can use the [Trend] transformation to remove linear trend from the series. This will not reduce quadratic or other curvilinear trend, however. A better method is to [Difference] the data. This transformation replaces values by the differences between each value and the previous value, thereby removing trend. For cyclical series, like monthly sales, seasonal differencing may be required before fitting a model (see below). Data that are drifting up or down across the series generally should be differenced.
2) **The variance of the series is constant across time.** If the series variation is increasing around its mean level across time, try a LOG transformation. If it is decreasing around its mean level across time (a rare occurrence), try a SQUARE transformation. You should generally do this before differencing.
3) **The autocorrelations of the series depend only on difference in time points and not on the time period itself.** If the first half of the ACF looks different from the second, try seasonal differencing after identifying a period on which the data are fluctuating. Monthly, quarterly, seasonal, annual data often cycle this way.

ACF plots

The autocorrelation function plot displays the pattern of autocorrelations. We have seen in this introduction an ACF plot of the residuals from a linear fit to birth rate. The slow decay of the autocorrelations after the first indicates autoregressive behavior in the residuals.

© 1989, SYSTAT, Inc.

PACF plots

The partial autocorrelation function plot displays autocorrelations, but each one below the first is conditioned on the previous autocorrelation. The PACF shows the relationship of points in a series to preceding points after partialing out the influence of intervening points. We examine them for effects that do not depend linearly on previous (smaller lag) autocorrelations.

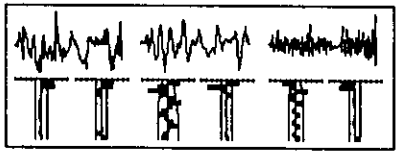
Identification using ACF and PACF plots together

Let's summarize our identification strategy. First, make sure the series is stationary. If variance is nonconstant, transform it with [Log] or [Square]. If trend is present, remove it with differencing. Finally, if seasonality is present, remove it with seasonal differencing. Then examine ACF and PACF plots together. Following are diagrams of possible types of patterns. Underneath each series is the ACF plot on the left and the PACF plot on the right. Several rows contain more than one possible plot. This is because coefficients in the model can be negative or positive and these plots show different combinations of signs.

Finally, remember that differencing can remove both trend and autoregressive effects. If an AR(1) model fits your data, as with our birth rate example, then differencing will produce only white noise and your ACF will look uniformly random. As a result, differencing is like constraining an autoregressive parameter to be exactly one.

AR(1)

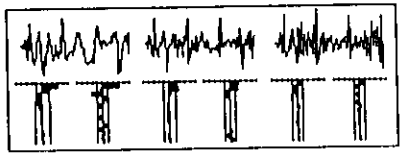
AR(2)



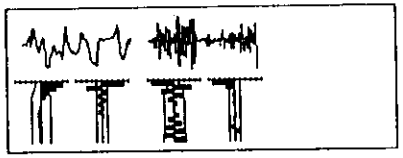
© 1989, SYSTAT, Inc.

MA(1)

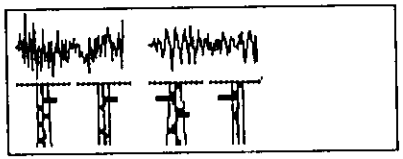
MA(2)



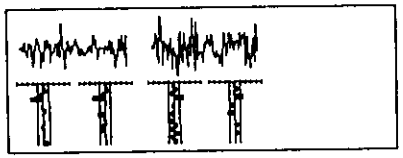
ARMA(1,1)



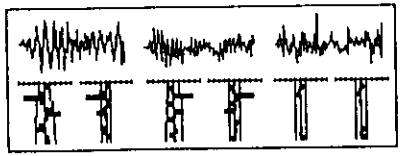
ARMA(0,0)(1,0)



ARMA(0,0)(0,1)



ARMA(0,0)(1,1)



© 1989, SYSTAT, Inc.

Estimating the ARIMA model

When you have identified the model as AR, MA, or a mixture, then you can fit it by specifying the AR order (P=*n*) and the MA order (Q=*n*). The "I" in ARIMA stands for "Integrated" and is a parameter that has to do with differencing. Any differencing you do while identifying the model will be included automatically in calculating your forecasts.

When you have estimated the model, pay attention to the standard errors of the parameters. If a parameter estimate is much smaller in absolute value than two standard errors away from zero, then it is probably unnecessary in the model. Refit the model without it. If you are uncertain about model identification, you can sometimes use this rule of thumb to compare two different models. The *mean square error* (MSE) of the model fit can also guide you. Generally, you are looking for a parsimonious model with small MSE.

Problems with forecast models

Forecasting is a vast field, and we cannot begin to explain even the basics in a chapter introduction. Makridakis, Wheelwright and McGee (1983) cover the topic fairly extensively. SYSTAT contains several methods for forecasting, with which you can experiment on these data. Exponential smoothing, for example, should provide similar forecasts to the ARIMA model. Keep in mind several things as you go:

1) There is nothing like extrinsic knowledge. We use forecasting methods at SYSTAT for budget planning. We always compare them to staff predictions of sales, however. In general, averaging staff predictions does better than the data-driven forecasting models. The reason is simple: staff know about external factors that are likely to affect sales. These are one-time events that are not easily included in models. Although we are not experts on the stock market, we would bet the same is true for investing. "Chartist" models that are based solely on the trends in stocks will not do as well, on average, as strategies based on knowledge of companies' economic performance and, in the illegal extreme, inside trading information.

2) Always examine your residuals. The same reasons for using residual diagnostics in ordinary linear regression apply to nonlinear forecasting models. In both cases, you want to see independence, or *white noise.*

3) Don't extrapolate too far. As in regression, predictions beyond the data are shaky. The farther you stray from the ends of the data, the less reliable are the predictions. The confidence limits on the forecasts will give you some flavor of this.

© 1989, SYSTAT, Inc.

Box and Jenkins (1976) provide the primary reference for these procedures. Financial forecasters should consult Nelson (1973) and Vandaele (1983) for applied introductions. Social scientists should look at McCleary and Hay (1980) for applications to behavioral data.

Many treatments (including Box and Jenkins) outline the ARIMA modeling process in three stages: Identification, Estimation, and Diagnosis. This is the outline we have followed in this introduction. With SYSTAT you identify models with **Transform...** ,**Case plot...**, **ACF plot...**, and **PACF plot...**, estimate them with **ARIMA...**, and diagnose their adequacy with more plots. For more complex problems, you may have to use other procedures, also.

ARIMA—AutoRegressive Integrated Moving Average—models can fit many time series with remarkably few parameters. Sometimes, ARIMA and Fourier models can be used effectively on the same data. As with other modeling procedures, decisions about appropriateness of competing models must rest on theoretical grounds. Nevertheless, a researcher should lean toward ARIMA models when it is reasonable to assume that points in a process are primarily functions of previous points and their errors, rather than periodic signal plus noise.

Seasonal decomposition and adjustment

A time series can be viewed as a sum of individual components that may include a term for location (level or mean value), a trend component (long term movements in the level of a series over time), a seasonal component, and an irregular component (the part unique to each time point). We can use the [Mean] transformation to remove the mean (location) from a series, [Trend] to remove a linear trend from a series, and [Difference] to eliminate either a trend or a seasonal effect from a series. Each of these transformations changes the scale of the series but does not directly provide information about the form of the trend or the seasonal component.

Alternatively, you may want to adjust the values in a series for the seasonal component but leave the series in the same scale or unit. This enables you to interpret the value units in the same way as the original series and to compare values in the series after removing differences due to seasonality.

© 1989, SYSTAT, Inc.

For example, sales data for many products are strongly seasonal. More suntan lotion is sold in the summer than in the winter. It is therefore difficult to compare suntan lotion sales from month to month (going up? going down?) without first taking seasonal differences into account.

Seasonal differences can be accounted for by determining a factor for each period of the cycle. Quarterly data may have a seasonal factor for each of the four quarters. Monthly data may have a seasonal factor for each of the twelve months.

Seasonal factors can take either of two forms: additive (fixed) or multiplicative (proportional). An additive seasonal factor is a fixed number of units above or below the general level of the series; e.g., 10,000 more bottles of suntan lotion were sold in July than the average month. In a multiplicative or proportional model, the seasonal factor is a percentage of the level of the series; e.g., 200% more bottles of suntan lotion were sold in July than in the average month.

Additive seasonal effects are removed from a series by subtracting estimates of the appropriate seasonal factor from each point in the series. Multiplicative seasonal effects are removed by dividing each point by the appropriate seasonal factor. **Seasonal...** computes either additive or multiplicative seasonal factors for a series and uses them to adjust the original series.

Exponential smoothing

Exponential smoothing forecasts future observations as weighted averages (a running smoother) of previous observations. For simple exponential smoothing, each forecast is the new estimate of location for the series. For models with trend and/or seasonal components, exponential smoothing smooths the location, trend, and seasonal components separately. For each component you must specify a smoothing weight between zero and one. In practice weights between .10 and .30 are most frequently used.

The [Exponential smoothing] option allows you to specify a linear or percentage growth (also called exponential or multiplicative) trend or neither, and an additive or multiplicative seasonal component or neither. There is always a location component. Thus, there are nine possible smoothing models from which you can choose.

© 1989, SYSTAT, Inc.

Smoothing with a linear trend component and no seasonal component is Holt's method. Smoothing with both a linear trend and a multiplicative seasonal term is Winter's three-parameter model.

The exponential smoothing procedure obtains initial estimates of seasonal components in the same manner as **Seasonal...**. If there is a trend component, SYSTAT uses regression (after adjusting values for any seasonal effects) to estimate the initial values of the location and trend parameters. If there is neither a trend nor a seasonal component, the first value in the series is used as the initial estimate of location.

Fourier analysis

If you believe your series is cyclical—such as astronomical or behavioral data—then you should consider Fourier analysis. The Fourier model decomposes a series into a finite sum of trigonometric components—sine and cosine waves of different frequencies. If your data are cyclical at a particular frequency, such as monthly, then a few Fourier components might be sufficient to capture most of the nonrandom variation.

Fourier analysis decomposes a time series just as a musical waveform can be decomposed into a fundamental wave plus harmonics. The French mathematician Fourier devised this decomposition around the beginning of the nineteenth century and applied it to heat transfer and other physical and mathematical problems. This transformation is of the general form:

$$f(t) = x_1 + x_2\sin(t) + x_3\cos(t) + x_4\sin(2t) + x_5\cos(2t) + \ldots$$

The Fourier decomposition can be useful for designing a filter to smooth noise and for analyzing the spectral composition of a time series. The most frequent application involves constructing a periodogram, which displays the squared amplitude (magnitude) of the trigonometric components versus their frequencies. **Fourier...** can be used to construct these displays. For further details on Fourier analysis, consult Brigham (1974) or Bloomfield (1976).

Fourier transforms are time consuming to compute because they involve numerous trigonometric functions. Cooley and Tukey (1965) developed a fast algorithm for computing the transform on a discrete series that makes the spectral analysis of lengthy series practical. A variant of this Fast Fourier Transform algorithm is implemented in SYSTAT.

© 1989, SYSTAT, Inc.

The discrete Fourier transform should be done on series with lengths (number of cases) that are powers of 2. If you do not have samples of 32, 64, 128, 256, etc., you should pad your series with zeros up to the next power of 2. If you have a series called Series with only 102 cases, for example, you can use the **Editor** menu to add zeros to cases 103 through 128. If you do not pad the file in this way, the Fourier procedure finds the highest power of 2 less than the number of cases in the file and transforms only that number of cases. (In this example, it would have transformed only the first 64 cases.)

A useful graph to accompany Fourier analysis is the *periodogram*. This graph plots magnitude (or squared magnitude) against frequency. It reveals the relative contribution of different frequency waveforms to the overall shape of the series. If the periodogram contains one large spike, as in the Newark temperature example later in this chapter, then it means that the series can be fit well by a single sinusoidal waveform. The periodogram is itself like a series, so sometimes you may want to smooth it with one of the **Series** smoothers

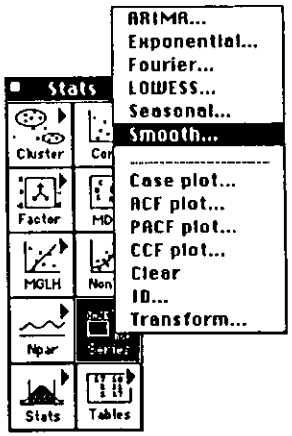
Fourier analysis is often used to construct a *filter*, which works like running smoothers. A filter allows variation of only a limited band of frequencies to pass through. A low-pass filter, for example, removes high frequency information. It is often used to remove noise in radio transmissions, recordings, and photographs. A high-pass filter, on the other hand, removes low frequency variation. It is used as one method for detecting edges in photographs. You can construct filters in SYSTAT by computing the Fourier transform, deleting real and imaginary components for low or high frequencies, and then using the inverse transform to produce a smoothed waveform.

Take a look at the monthly temperature data example (9.8) to see how Fourier analysis identifies frequency components. If you reproduce a series from a few low frequency Fourier components, the resulting smooth will be similar to that achieved by a running window of an appropriate width. The Fourier method will constrain the smooth to be more regularly periodic, however, since the selected trigonometric components will completely determine the periodicity of the smooth.

© 1989, SYSTAT, Inc.

Usage

Series lets you analyze series data interactively. The general strategy is to plot the series using **Case plot...**, **ACF plot...**, **CCF plot...**, or **PACF plot...**; smooth or transform the series using **ARIMA...**, **Exponential...**, **Fourier...**, **LOWESS...**, **Seasonal...**, **Smooth...**, and **Transform...**; and examine the results by plotting the smoothed or forecasted series.



To use an item, press and hold the mouse button on **Series**. A submenu pops up offering **ARIMA...**, **Exponential...**, etc. Select the item you want and release the mouse button.

By default, **Series** works in working memory of the computer (RAM), never altering the data file containing a series. You may, if you wish, save the results of any transformation to a SYSTAT file by clicking the [Save file] option in the dialog box. Generally, you will want to experiment in memory and then save to a file when you have results that please you. We discuss the **Series** items in the order you're likely to use them:

© 1989, SYSTAT, Inc.

Utilities	**Transform...** **ID...** **Clear**
Plots	**Case plot...** **ACF plot...** **PACF plot...** **CCF plot...**
Models	**ARIMA...** **Smooth...** **LOWESS...** **Exponential...** **Seasonal...** **Fourier...**

Transform...

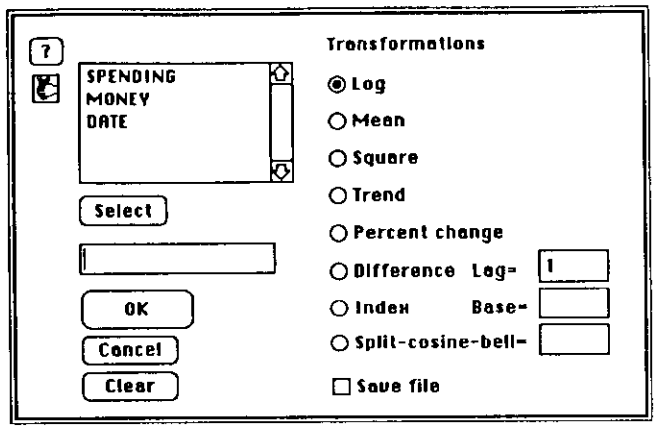
Transform... provides log, mean, square, difference, percent, index, trend, and split-cosine-bell taper transformations.

- Select **Series/Transform...** from the **Stats** menu
- Select a variable to be transformed; the first numeric variable is used by default
- Choose the type of transformation to be performed
  - Specify a [Lag] for the [Difference] transformation
  - Specify a [Base] value for the [Index] transformation; the default is the first value of the series
  - Specify a proportion between 0 and 1 for the [Split-cosine-bell] taper
- Click OK

You can pile up transformation commands in any order, as long as you do not encounter a mathematically undefined result. In that case, SYSTAT gives an error message and the variable is restored to its original value in the file.

Several of the examples illustrate these transformations. In identifying ARIMA models, for example, we sometimes log the series, look at an ACF plot, difference the series, look at another plot, and so on. **Series** allows you to do plots and transformations as you go so you don't have to commit yourself at once to complicated sets of transformations.

© 1989, SYSTAT, Inc.



Log	[Log] replaces the values in a series with their natural logarithms. It removes nonstationary variability, such as increasing variance across time.
Mean	[Mean] subtracts the mean from each value in a series.
Square	You can also [Square] the values in a series. This is useful for producing periodograms for spectral analysis (see Fourier section below) and for normalizing variance in series.
Trend	[Trend] removes linear trend from a series. Fourier analysis, for example, should generally begin with a "level" series.
Percent change	This replaces each value in a series by the difference from the previous value expressed as a percentage change—the difference in values divided by the previous value.
Difference	An important transformation for many series is the differencing operator. [Difference] replaces values by the differences between each value and the previous value, thereby removing trend (nonstationarity in level over time). The first point in the series is set to a missing value since there is no preceding point to subtract from it. This transformation is useful to make some series "stationary." If each value in a series depends on the preceding point's value, then differencing removes this dependence. If the dependence is perfect, the resulting series will be all zeroes. Usually, however, the remaining variation is random.

© 1989, SYSTAT, Inc.

Sometimes you may want to remove seasonal dependence. For example, in an economic series, January production may be similar to the production from the previous January. In this case, you can subtract the first point in the series from the twelfth, the second from the thirteenth, etc., by typing 12 for [Lag].

The default value for [Lag] is 1. You cannot specify a lag that is longer than the series, but you can difference a series several times. For example, you could difference it first with the default lag of 1, and then difference it again with lag 12. This leaves the first 13 values of the series missing after taking out first order and seasonal lags of 12. If the differencing is done in the reverse order, the result is the same. The interested algebra student is left to prove why this is true (you're not interested?).

Index

[Index] replaces each value in a series by a ratio of the value to the value of a base observation. If you do not specify a base, [Index] uses the value of the first observation as the base value. A value of 2 in the transformed series means that the original value was twice that of the base value; a value of 0.5 means the value was one-half the base.

Split-cosine-bell

The split-cosine-bell taper is used prior to a Fourier decomposition to reduce "leakage" between components. Tapering weights the middle of a series more than the endpoints. It is a general smoother.

Specify the proportion (P) of the series to be tapered in the box. You can choose a weighting function that varies between a "boxcar" (P=0) and a full cycle of a cosine wave from trough to trough (P=1). For intermediate values of P, the weighting function is flat in the center section and cosine tapered at either end. If you do not specify a P value, it is assumed to be .5.

Saving transformed series

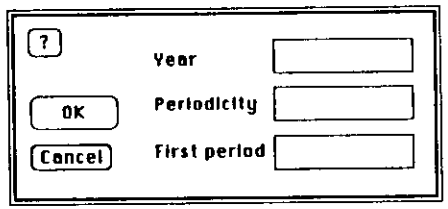
If you click [Save file], the transformed series is saved into a SYSTAT file. The name of the series is changed to TRANSF.

© 1989, SYSTAT, Inc.

ID...

Use **ID** to label the cases of a variable in a series plot. You specify up to three values, [Year], [Periodicity], and [First period].

- Select **Series/ID...** from the **Stats** menu
- Specify values for [Year], [Periodicity], and [First period]
- Click OK



To indicate a monthly series (12 periods per seasonal cycle) starting in January of 1952, type 1952 for [Year], 12 for [Period], and 1 for [First period].

If the series starts in March of that year (the third period), type 3 in the [First period] box.

ID... labels the values of a series with year and period. For example, you can label quarterly data by specifying the first year (e.g., 1952), the periodicity (e.g., 4 quarters per year), and the period of the first observation (e.g., the 4th quarter of 1952).

Clear

Clear restores the original values of all series. If you have transformed or smoothed several series, it clears them all. **Clear** does not affect any transformations or smoothings saved in SYSTAT files.

- Select **Clear** from the **Stats** menu

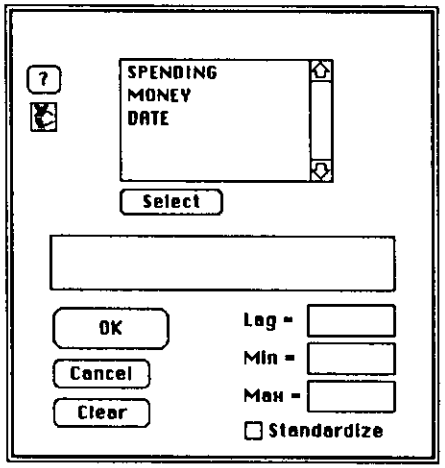
Clear clears the series transformation from memory and restores the original values of the series (the values that are still sitting unchanged in the Data Editor). It is not possible to clear only the latest transformation (unless you are saving to files after each step)—**Clear** undoes ***all*** the transformations.

© 1989, SYSTAT, Inc.

Case plot...

Case plot... provides time series plots. The variable you select is the dependent (vertical axis) variable, and the Case is the independent variable. The points are connected with a line.

- Select **Series/Case plot...** from the **Stats** menu
- Select a variable to be analyzed; the default is the first numeric variable(s) in the file
- Click OK



By default, SYSTAT plots all cases; use [Lag] to plot fewer. If the value for [Lag] is greater than or equal to the number of cases in the file, all the cases are plotted.

You can control the extreme values of the Y (vertical) axis with [Min] and [Max]. For example, to have a vertical scale of 10 to 75, set [Min] to 10 and [Max] to 75.

[Standardize]

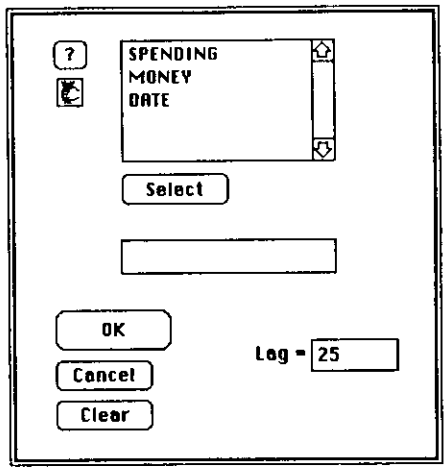
Click [Standardize] to standardize the variable before plotting. Standardizing subtracts the sample mean from each case and divides the difference by the sample standard deviation. Standardizing is useful for overlaying plots of data with different scales.

© 1989, SYSTAT, Inc.

ACF plot...

ACF plot... draws autocorrelation plots.

- Select **Series/ACF plot...** from the **Stats** menu
- Select a variable to be analyzed; the default is the first numeric variable in the file
- Click OK



By default, SYSTAT plots the first 25 cases; use [Lag] to plot more or fewer. If the value for [Lag] is greater than or equal to the number of cases in the file, all the cases are plotted.

PACF plot...

PACF plot... draws partial autocorrelation plots. Its dialog box is the same as that of **ACF plot...** and **CCF plot...**

- Select **Series/PACF plot...** from the **Stats** menu
- Select a variable to be analyzed; the first numeric variable in the file is used by default
- Click OK

© 1989, SYSTAT, Inc.

CCF plot...

CCF plot... draws cross-correlation plots. Its dialog box is the same as that of **ACF plot...**

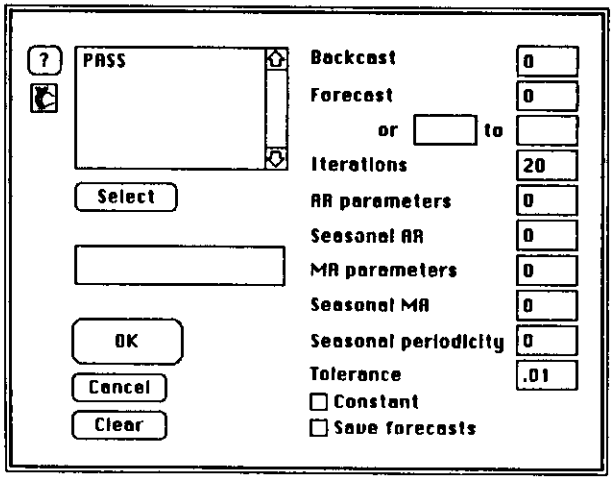
- Select **Series/CCF plot...** from the **Stats** menu
- Select two variables to be analyzed; the first numeric variables in the file are used by default
- Click OK

Cross-correlation plots help identify relations between two different series and any time delays to the relations. A correlation for a negative lag indicates the relation of the values in the first series to values in the second series that number of periods earlier. The correlation at lag zero is the usual contemporaneous Pearson correlation. Similarly, correlations at positive lags relate values in the first series to subsequent values in the second series.

ARIMA...

ARIMA... provides ARIMA (AutoRegressive Integrated Moving Average) models for time series.

- Select **Series/ARIMA...** from the **Stats** menu
- Select the variable to be analyzed; the default is the first numeric variable
- Click [Constant] to add a constant term (sample mean) to the model
- Click OK



© 1989, SYSTAT, Inc.

Parameters

You can specify the number of AR and MA parameters with the [AR parameters] and [MA parameters] options. If you specify more than one AR or MA parameter, SYSTAT automatically includes all parameters up to that order.

Seasonal models

For seasonal models, specify the number of parameters with the [Seasonal AR] and [Seasonal MA] options. The default periodicity is 12; you can specify a different [Periodicity]. This periodicity applies to all seasonal AR and MA parameters.

ARIMA constant

You can estimate a constant in your model by clicking the [Constant] box. If you have differenced your data, this is usually unnecessary.

Iterations and tolerance

You can control iterations by specifying a maximum number of [Iterations] and/or a convergence criterion [Tolerance]. Don't play around with tolerance unless you are failing to get convergence of the estimates after many iterations. It is better to increase the number of iterations than to decrease the tolerance criterion, since your estimates will be more precise. In any case, [Tolerance] cannot be set greater than a tenth. Sometimes models fail to converge after many iterations because you have misspecified them.

Forecasting and backcasting

Specify the number or range of cases you want forecasted with the [Forecast] option.

You can backcast (forecast backwards—i.e., estimate what "previous" cases were). Just specify the number of cases you want backcasted with the [Backcast] option. This often improves MA and seasonal MA estimates. Choose a number for backcasting that is at least 10 or so or greater than your seasonal period.

Saving forecasts

You can save forecasts into a SYSTAT file with the [Save forecasts] option. The forecasts and residuals are saved with the variable names FORECAST and RESIDUAL. Actually, the forecasted values are appended to the series, so the file contains original values plus forecasts. For the airline data in Example 9.5, the saved file looks like this:

x = original data (not logged or differenced)
r = residual (for logged and differenced model)
f = forecast (in original data units)
. = missing value

© 1989, SYSTAT, Inc.

CASE	FORECAST	RESIDUAL
1	x	.
2	x	.
3	x	.
4	x	.
5	x	.
6	x	.
7	x	.
8	x	.
9	x	.
10	x	.
11	x	.
12	x	.
13	x	.
14	x	r
15	x	r
16	x	r
	and so on	
144	x	r
145	f	.
146	f	.
147	f	.
148	f	.
149	f	.
150	f	.
151	f	.
152	f	.
153	f	.
154	f	.

Notice that there are no residuals for the first 13 points because of the differencing (previous commands) or for the final 10 forecasts. Remember also that the residuals are for the logged model so they are not in the same units as the raw data and forecasts if you log a series.

If you forecast somewhere in the middle of a series (by specifying a range of cases), the forecasts (*f*) are written over the original series values (*x*) in those locations.

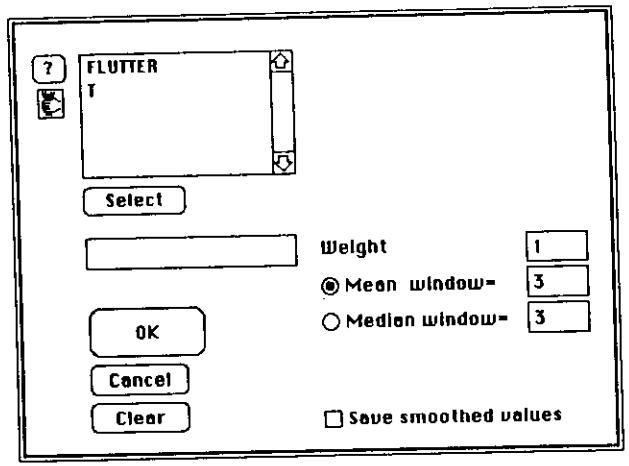
If you are satisfied with your ARIMA model and want to plot the forecasts, you can open the file in which you saved the forecasts and plot them. Remember, however, that this cleans out your active work area.

© 1989, SYSTAT, Inc.

Smooth...

Series/Smooth... provides a variety of methods for smoothing time series. Running means (moving averages), running medians, and general linear filters are all possible.

- Select **Series/Smooth...** from the **Stats** menu
  - Select a variable; the first numeric variable is used by default
  - Choose mean or median smoothing and specify a window size
  - Specify [Weight] values for weighted smoothing
- Click OK



Smooth... transforms the weights before using them so that their sum is 1; e.g., 1,2,1 is equivalent to setting Weight to .25,.5,.25 or 3,6,3.

Saving smoothed values

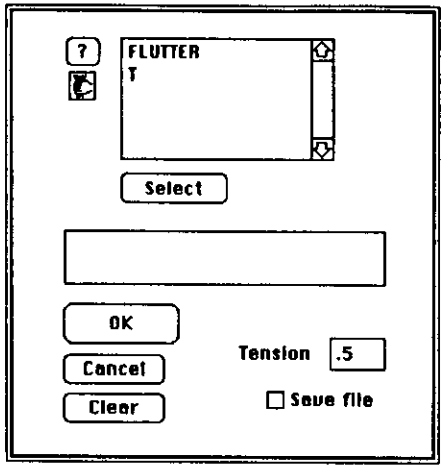
You can save smoothed values into a SYSTAT file with the [Save smoothed values] option. The series is renamed SMOOTH.

LOWESS...

LOWESS... provides Cleveland's LOWESS smoother.

- Select Series/LOWESS... from the Stats menu
  - Select a variable to be smoothed; the first numeric variable in the file is used by default
  - Specify a smoothness factor, or [Tension], between 0 and 1
- Click OK

© 1989, SYSTAT, Inc.



LOWESS is most often used for smoothing scatterplots. For scatterplot smoothing with LOWESS, see SYGRAPH. **Series...** assumes the "X" variable is equally spaced and the "Y" variable is the series. The [Tension] factor, called *F* by Cleveland, determines the stiffness of the smooth. It is related to window width in running smoothers. It varies between 0 and 1, with a default of .5.

Saving smoothed values

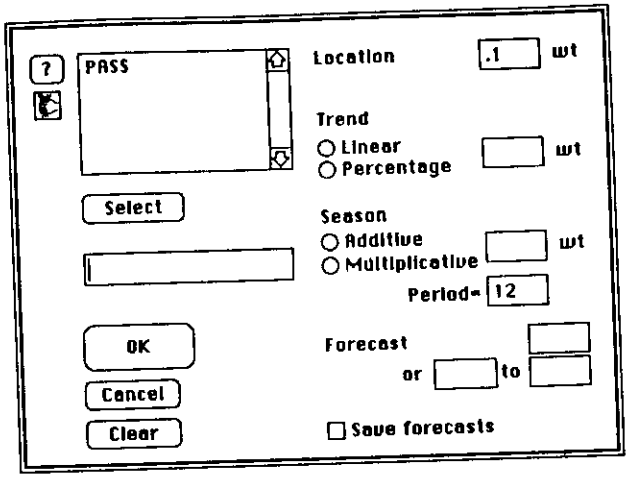
You can save smoothed values into a SYSTAT file with the [Save smoothed values] option. The series is renamed SMOOTH.

Exponential...

Series/Exponential... does exponential smoothing models with optional additive and multiplicative seasonal terms and linear and percentage growth.

- Select **Series/Exponential...** from the **Stats** menu
- Select a variable to be smoothed; the first numeric variable is used by default
- Choose any combination of location, ([Smooth]), trend ([Linear] or [Percentage]), and seasonal ([Additive] or [Multiplicative]) components and specify a weight to be used for each component
- Click OK

© 1989, SYSTAT, Inc.



The [Exponential smoothing] option allows you to specify a linear or percentage growth (also called exponential or multiplicative) trend or neither, and an additive or multiplicative seasonal component or neither. There is always a location component. Thus, there are nine possible smoothing models you can choose from. Smoothing with a linear trend component and no seasonal component is Holt's method. Smoothing with both a linear trend and a multiplicative seasonal term is Winter's three-parameter model.

The exponential smoothing procedure obtains initial estimates of seasonal components in the same manner as **Seasonal...**. If there is a trend component, SYSTAT uses regression (after adjusting values for any seasonal effects) to estimate the initial values of the location and trend parameters. If there is neither a trend nor a seasonal component, the first value in the series is used as the initial estimate of location.

Smoothing weight

Specify a smoothing weight in the [Smooth] box. This usually varies between .1 and .3. The default is .2.

© 1989, SYSTAT, Inc.

Additive or multiplicative seasonal models

For seasonal components, choose either an [Additive] or [Multiplicative] model. The default is neither. You may supply a weight in the accompanying box if you choose either option. This weight is usually between .1 and .3. The default is .2. You may use any combination of the [Additive] or [Multiplicative] choices with the [Linear] or [Percentage] choices of growth model. Specify periodicity in the [Period] box; the default is 12.

Linear or percentage models

Select either a [Linear] or [Percentage] model. Percentage growth models are sometimes called exponential or multiplicative growth models. The default is neither selection.You may supply a weight in the accompanying box if you choose either option. This weight is usually between .1 and .3. The default is .2. You may use any combination of the [Linear] or [Percentage] choices with the [Additive] or [Multiplicative] seasonal component choices.

Forecasting

You can predict new cases by specifying the number or range of cases for [Forecast].

Saving forecasts

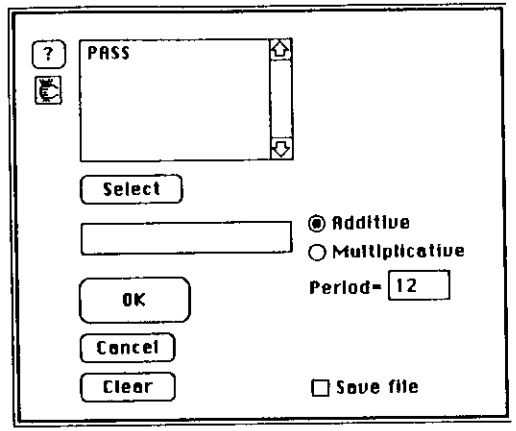
You can save forecasts into a SYSTAT file with the [Save forecasts] option. The forecasts and residuals are saved with the variable names FORECAST and RESIDUAL.

Seasonal...

Seasonal... provides seasonal decomposition models, which remove additive or multiplicative effects.

- Select **Series/Seasonal...** from the **Stats** menu
- Select a variable; the first numeric variable is used by default
- Choose [Additive] or [Multiplicative] effects
- Click OK

© 1989, SYSTAT, Inc.



Seasonal factors can take either of two forms: additive (fixed) or multiplicative (proportional). An additive seasonal factor is a fixed number of units above or below the general level of the series. In a multiplicative or proportional model, the seasonal factor is a percentage of the level of the series. Additive seasonal effects are removed from a series by subtracting estimates of the appropriate seasonal factor from each point in the series. Multiplicative seasonal effects are removed by dividing each point by the appropriate seasonal factor. **Seasonal...** computes either additive or multiplicative seasonal factors for a series and uses them to adjust the original series.

Additive or multiplicative models

Click [Additive] to fit an additive model or click [Multiplicative] to fit an multiplicative model. Specify periodicity in the [Period] box; the default is 12.

Saving smoothed values

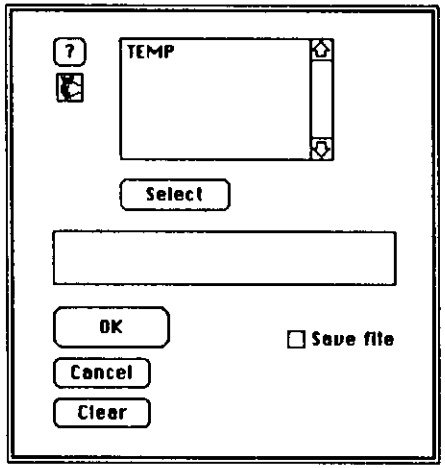
You can save deseasonalized values into a SYSTAT file with the [Save file] option. The series is renamed TRANSF.

© 1989, SYSTAT, Inc.

Fourier...

Fourier... computes Fourier decomposition.

- Select **Series/Fourier...** from the **Stats** menu
  - Select one or two variables to be analyzed; the first numeric variable is used by default
  - Click [Save file] to save the results of the transformation in a SYSTAT file
- Click OK



If you select two variables, SYSTAT assumes you want the inverse transformation. The first variable selected is used as the real component and the second variable is the imaginary.

Warning: the Fourier transformation is an "in place" transformation. The *magnitude* of the transformation is written over the active work area where you stored the variable you transformed. You can plot this magnitude with the **Series/Case plot...** item or even smooth it. If you want to save the results of the transformation, click [Save file].

Saving files

SYSTAT saves the real and imaginary components in a SYSTAT file. You can then manipulate those components with DATA. Real and imaginary components are saved instead of magnitude and phase because that allows you to do an inverse Fourier transform.

© 1989, SYSTAT, Inc.

Inverse Fourier

Here is how to do an inverse transform. Assume you have saved the results of a direct transform into a file MYFOUR. That file should contain two variables—REAL and IMAG—which are the two components of the transform.

- Select **Series/Transform...** from the **Stats** menu
- Choose the type of transformation to be performed: [Fourier]
- Select the variables to be analyzed: REAL, IMAG
- Click OK

Since you specify two variables, SYSTAT assumes you want the inverse transform, and that the first variable is the real component and the second is the imaginary. The work is done in the active work area, so the resulting real series is stored in the active work area occupied by REAL (or whatever you called the first variable corresponding to the real component).

Saving magnitude and phase

If you absolutely must have magnitude and phase in a SYSTAT file instead of the real and imaginary components, you can use the **Editor** menu to do the conversion.

- Select **Math...** from the **Editor** menu
- Fill in the set statement:
  Set Magnitude to SQR(REAL*REAL + IMAG*IMAG)
- Click OK

- Select **Math...** from the **Editor** menu again
- Fill in the set statement:
  Set PHASE to ATN(IMAG/REAL)
- Click OK

© 1989, SYSTAT, Inc.

Computation

Missing data

The **Series** procedures handle missing data in several ways. First, when data are read into the active work area, all missing values in a series are removed. If you have a file containing several series, be careful. Missing data in each series determines the sequence of nonmissing values in the active work area. Using several series with different patterns of missing data in the same SYSTAT file causes a different alignment for each one.

When values are copied back from the active work area into a SYSTAT file, missing values are inserted where they were in the original file series. This way, they are correctly aligned with the original data if you merge files. Again, each series determines where the missing values are inserted. Thus, it is best to keep series with different patterns of missing data in separate files.

As we have seen, missing data are generated for certain residuals when saved into files containing forecasts. See the Usage section above under **ARIMA...**, Saving forecasts.

Missing values in the active work area (caused by differencing) are ignored by smoothing, Fourier, and ARIMA procedures. You will see them when you plot data, however.

Algorithms

The LOWESS algorithm for XY and scatterplot smoothing is documented in Cleveland (1979) and Cleveland (1981). All the smoothers are computed in single precision.

The Fast Fourier Transform is due to Gentleman and Sande (1966), and documented further in Bloomfield (1976). The Fourier routines are in single precision.

© 1989, SYSTAT, Inc.

ARIMA models are estimated with a set of algorithms. Residuals and unconditional sums of squares for the seasonal multiplicative model are calculated by an algorithm in McLeod and Sales (1983). The sums of squares are minimized iteratively by a quasi-Newton method due to Fletcher (1972). A penalty function for inadmissible values of the parameters makes this procedure relatively robust when values are near the circumference of the unit circle. Standard errors for the parameter estimates are computed from the inverse of the numeric estimate of the Hessian matrix, following Fisher (1922). Forecasting is performed via the difference equations documented in Chapter 5 of Box and Jenkins (1976). Estimation is performed in double precision and forecasting is done in single precision.

© 1989, SYSTAT, Inc.

Examples

9.1 Series plots and differencing

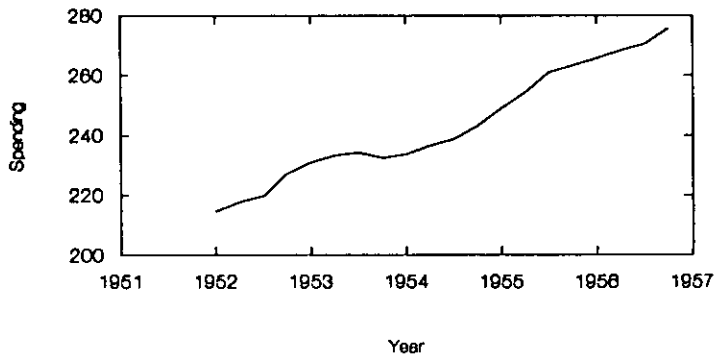
This example uses the SPNDMONY file. (If you have used the file already, be sure to **Clear**... the series before trying this example.) First, we label the series values:

- Select **Series/ID**... from the **Stats** menu
- Set [Year] to 1952, [Periodicity] to 4, and [First period] to 1
- Click OK

```
Series originates at: 1952  Periodicity:  4  First Period:  1
```

Now, we draw a time series plot.

- Select **Series/Case plot**... from the **Stats** menu
- Select the variable to be plotted: SPENDING
- Click OK



© 1989, SYSTAT, Inc.

When you opened SPNDMONY, the file was opened for reading just as with other SYSTAT procedures. However, unlike other procedures, when you drew a plot, **Series** read the values for SPENDING from the file and saved them into an "active work area." This area occupies space in the memory of the computer. Then the plot was constructed by reading the values from this active work area. If you plotted again, there would be no message about copying SPENDING to the work area. The SYSTAT file SPNDMONY would not be reread. Instead, the values would be read directly from the active work area in memory.

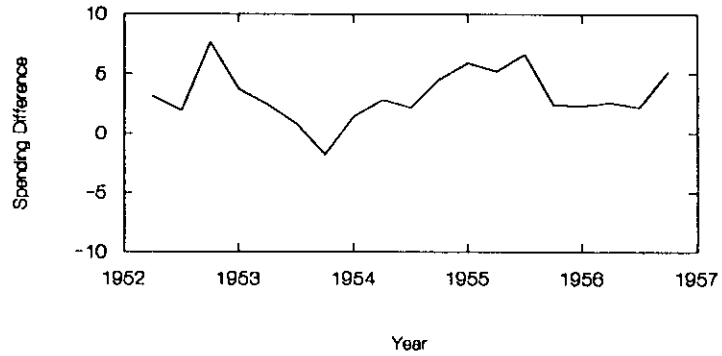
Now, we difference the series:

- Select **Series/Transform**... from the **Stats** menu
- Choose the type of transformation to be performed: [Difference]
- Select the variable to be transformed: SPENDING
- Click OK

```
Series is transformed
```

Now, your active work area contains differenced (detrended) values of SPENDING. If you make any further references to SPENDING in later commands, these new values are used instead of the original ones in the file. Now a series plot shows the differenced series:

- Select **Series/Case plot**... from the **Stats** menu
- Select the variable to be plotted: SPENDING
- Click OK

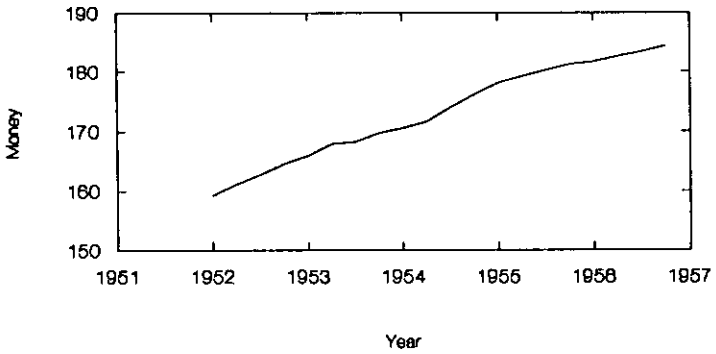


© 1989, SYSTAT, Inc.

The strong increasing trend of the values in the series has been removed.

Now we plot the MONEY series.

- Select **Series/Case plot...** from the **Stats** menu
- Select the variable to be plotted: MONEY
- Click OK



The values for MONEY are in the SYSTAT file, but not in your active work area. Consequently, before the plot is constructed, these values are read into your active work area. You are always notified when variables are read from a SYSTAT file into your active work area. If you do not get a message about this, you can assume your command is operating directly on the variable in the active work area. At this time both SPENDING and MONEY are in the active work area.

As you can imagine, if you try to plot too many variables, you will run out of space in the active work area (memory in the machine). **Series** is not intended for plotting many variables. Use SYGRAPH for that. Use **Series** to work interactively with one or two series at a time. To reclaim space in the work area, use **Clear**.

Clear removes the series MONEY from memory, providing room for another series.

© 1989, SYSTAT, Inc.

What if you want to work with the raw data for SPENDING again? Remember, we differenced SPENDING, so it is no longer the same. The answer is to use **Clear**. **Clear** clears the active work area filled by SPENDING. The next time you use SPENDING in a command, the SPENDING values are read from the file again.

Transformations and smoothings change only the values in your active work area. This helps you to explore various combinations of transformations and do plots to examine their effect.

We could use **Transform...** to [Log] the series. We could then use **Transform...** again to [Mean] the series.

The first transformation brings SPENDING into the active work area. Then each value is (natural) logged. Then the mean of the logged series is subtracted from each logged value.

What is different (and wrong) with doing a [Mean] transformation first and then a [Log] transformation?

The first transformation removes the mean from SPENDING as we did in the previous example. Then the logs are taken. Some of the values are negative, however, since the mean of the series was removed. Since we cannot take logs of negative numbers, SYSTAT returns an error message and sets the illegal values to missing. We must either clear SPNDMONY or open the file again and try a different sequence of transformations.

9.2 ACF, PACF, CCF plots

Autocorrelation, partial autocorrelation, and cross-correlation plots are useful for analyzing residuals, investigating relations within and between series, and identifying ARIMA models.

Autocorrelation plots

An autocorrelation is like an ordinary correlation of a series with itself shifted once, twice, etc. If each point is related to the previous point's value, then this correlation is high. The second autocorrelation is like the correlation of a series with itself shifted back twice. This second autocorrelation should reveal dependencies that go back two points.

© 1989, SYSTAT, Inc.

Many examples in this chapter use the SPNDMONY file, which contains two quarterly series, SPENDING (consumer expenditures) and MONEY (money stock) in billions of current dollars for the United States in the years 1952-1956. These series are analyzed by Chatterjee and Price (1977).

We can get the ACF plot of SPENDING:

- Select **Series/ACF plot...** from the **Stats** menu
- Select the variable to be plotted: SPENDING
- Click OK

```
Plot of SPENDING
Number of cases =   20
Mean of series =     243.460
Standard deviation of series =      10.273

Plot of Autocorrelations

 Lag   Corr     SE -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
                      |____|____|____|____|____|____|____|____|____|____|
   1   .839   .224                    (         |**********)*********
   2   .685   .347               (              |**************)
   3   .521   .409          (                   |*************       )
   4   .382   .441       (                      |*********             )
   5   .254   .457       (                      |******                )
   6   .131   .464    (                         |***                     )
   7   .026   .466    (                         |                        )
   8  -.077   .466    (                        *|                        )
   9  -.163   .467    (                    ****|                        )
  10  -.233   .470    (                  *****|                        )
  11  -.300   .475    (                *******|                        )
  12  -.348   .485  (                  ********|                          )
  13  -.371   .497  (                 *********|                          )
  14  -.391   .511  (                 *********|                          )
  15  -.386   .525  (                 *********|                          )
```

The autocorrelations are initially large and then become negative. This is typical of trended series. Values that are close together in the time sequence tend to be similar in magnitude (both large or both small) because of the trend. For the same reason values far apart tend to be opposite in magnitude.

To understand better the dynamics of the series, we can difference it to remove the trend and plot it again.

- Select **Series/Transform...** from the **Stats** menu
- Choose the type of transformation: [Difference]
- Select the variable to be transformed: SPENDING
- Click OK

© 1989, SYSTAT, Inc.

- Select **Series/ACF plot...** from the **Stats** menu
- Select the variable to be plotted: SPENDING
- Click OK

```
Plot of SPENDING
Number of cases =   19
Mean of series =       3.211
Standard deviation of series =       2.158

Plot of Autocorrelations

 Lag   Corr     SE -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
                      |____|____|____|____|____|____|____|____|____|____|
   1   .360   .229                    (         |********  )
   2   .117   .257                  (           |**          )
   3  -.143   .260                 (      ***    |              )
   4  -.477   .264                 ( ***********|              )
   5  -.219   .306              (     *****     |                )
   6  -.241   .314              (     ******    |                )
   7  -.232   .324              (     *****     |                 )
   8   .092   .333              (               |**               )
   9   .164   .334              (               |****             )
  10   .154   .338              (               |***              )
  11   .172   .342              (               |****              )
  12  -.178   .346              (          ****|                  )
  13  -.086   .351              (            **|                  )
  14  -.040   .352              (              |                   )
  15  -.029   .352              (              |                   )
```

There no longer seems to be any relation between values in the series and their previous values. For more information on these plots, see ARIMA below or the references cited in the introduction to this chapter.

© 1989, SYSTAT, Inc.

Partial autocorrelation plots

Partial autocorrelations are like ordinary partial correlations. Partial autocorrelation plots show the relationship of points in a series to preceding points after partialing out the influence of intervening points. We can examine them for effects that do not depend linearly on previous (smaller lag) autocorrelations.

Let's try a partial autocorrelation plot of SPENDING. **Clear** removes the differenced values from memory.

- Select **Series/Clear** from the **Stats** menu

- Select **Series/PACF plot...** from the **Stats** menu
- Select the variable to be plotted: SPENDING
- Click OK

```
Plot of SPENDING
Number of cases =   20
Mean of series =     243.460
Standard deviation of series =      18.273

Plot of Partial Autocorrelations
```

Lag	Corr	SE	-1.0 -.8 -.6 -.4 -.2 .0 .2 .4 .6 .8 1.0
1	.839	.224	(|*********)
2	-.065	.224	(* |)
3	-.122	.224	(*** |)
4	-.025	.224	(|)
5	-.058	.224	(* |)
6	-.094	.224	(** |)
7	-.041	.224	(* |)
8	-.097	.224	(** |)
9	-.063	.224	(* |)
10	-.050	.224	(* |)
11	-.100	.224	(** |)
12	-.056	.224	(* |)
13	-.019	.224	(|)
14	-.088	.224	(** |)
15	-.016	.224	(|)

The first autocorrelation is the same as in the ACF plot. There are no previous autocorrelations, so it is not adjusted. The second-order autocorrelation was .685 in the ACF plot, but after adjusting for the first autocorrelation it is reduced to –.065. The other higher-order autocorrelations have been affected similarly.

See the ARIMA section below for applications and check the references cited there.

Cross-correlation plots

Cross-correlation plots help identify relations between two different series and any time delays to the relations.

© 1989, SYSTAT, Inc.

A correlation for a negative lag indicates the relation of the values in the first series to values in the second series that number of periods earlier. The correlation at lag zero is the usual contemporaneous correlation. Similarly, correlations at positive lags relate values in the first series to subsequent values in the second series.

- Select **Series/CCF plot...** from the Stats menu
- Choose the type of plot to be drawn: CCF
- Select the variables to be plotted: SPENDING, MONEY
- Click OK

```
Plot of Cross Correlations
```

Lag	Corr	SE	-1.0 -.8 -.6 -.4 -.2 .0 .2 .4 .6 .8 1.0
-7	.147	.277	(|***)
-6	.259	.267	(|******)
-5	.368	.258	(|*********)
-4	.481	.250	(|************)
-3	.595	.243	(|************)**
-2	.731	.236	(|**********)*******
-1	.857	.229	(|**********)**********
0	.978	.224	(|**********)************
1	.803	.229	(|**********)********
2	.641	.236	(|**********)*****
3	.481	.243	(|************)
4	.332	.250	(|********)
5	.188	.258	(|****)
6	.060	.267	(| *)
7	-.055	.277	(* |)

There is strong contemporaneous correlation (at lag 0) between the two series, tapering off the further one goes in either direction. This is true of all cross-correlation functions between two trended series. Since both series are increasing, early values in both series tend to be small and final values tend to be large. This produces a large positive correlation.

© 1989, SYSTAT, Inc.

To better understand the relationship, if any, between the series, difference them to remove the common trend and then display a new CCF plot.

- Select **Series/Transform...** from the **Stats** menu
- Choose the type of transformation: [Difference]
- Select the variable to be transformed: SPENDING
- Click OK

- Select **Series/Transform...** from the **Stats** menu
- Choose the type of transformation: [Difference]
- Select the variable to be transformed: MONEY
- Click OK

- Select **Series/CCF plot...** from the **Stats** menu
- Select two variables: SPENDING, MONEY
- Click OK

```
Series is transformed

Plot of cross correlations
 Lag   Corr     SE -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
                    |____|____|____|____|____|____|____|____|____|____|
  -7  -.048   .289              (                .                )
  -6  -.395   .277              (   *********    |                )
  -5  -.405   .267              (   *********    |                )
  -4  -.123   .258                (        ***   |                )
  -3   .004   .250                (              |                )
  -2   .342   .243                (              |********        )
  -1   .368   .236                 (             |*********       )
   0   .250   .229                 (             |******          )
   1   .098   .236                 (             |**              )
   2  -.068   .243                 (                .             )
   3  -.617   .250               ***(***********    |             )
   4  -.494   .258                 (***********     |             )
   5  -.516   .267                (************    |               )
   6  -.085   .277                (               **|               )
   7   .265   .289                (                 |******         )
```

This shows a significant negative correlation at only one time interval, +3 lags of the series. Since we selected SPENDING first, we see that consumer expenditures are negatively correlated with the money stock three quarters later. Thus, consumer spending may be a "leading indicator" of money stock.

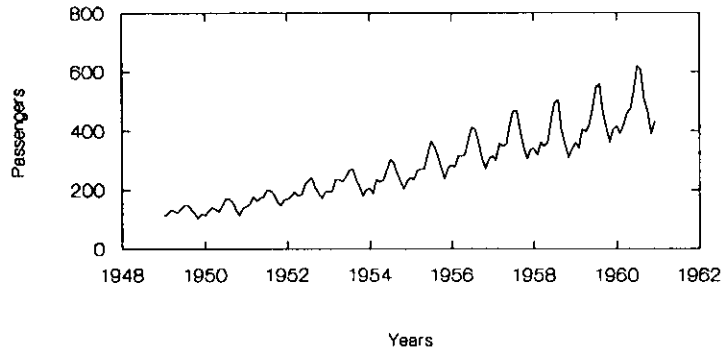
© 1989, SYSTAT, Inc.

9.3 ARIMA model identification

The following data (in the file AIRLINE) are monthly totals of international airline passengers for twelve years beginning January, 1949. Box and Jenkins (1976) call this Series G. We plot this series so you can see the strong periodicities. Notice that these cycles do not seem to be smooth functions of sine and cosine waves like the Newark temperature data in Example 9.8. Instead there are jagged spikes in the data that correspond, apparently, to holidays like Christmas and Easter.

- Select **Series/ID...** from the **Stats** menu
- Specify Year=1949, Periodicity=12
- Click OK

- Select **Series/Case plot...** from the **Stats** menu
- Specify the variable to be plotted: PASS
- Click OK



The first thing to worry about in our modeling is the increasing variance in the series over time. One way to deal with this is to log the data:

- Select **Series/Transform...** from the **Stats** menu
- Choose [Log]
- Select the variable to be transformed: PASS
- Click OK

Next, we should identify which ARIMA parameters we wish to estimate by plotting the data several ways.

© 1989, SYSTAT, Inc.

These are the parameters of the ARIMA model.

AR	*autoregressive*	each point is a weighted function of a previous point plus random error
I	*integrated*	each point's value is a constant difference from a previous point's value
MA	*moving average*	each point is a weighted function of a previous point's random error plus its own random error

For seasonal ARIMA models, we need three additional parameters: Seasonal AR, Seasonal I, and Seasonal MA. Their definitions are the same as above, except they apply to points that are not adjacent in a series. The AIRLINE data involve seasonal parameters, for example, because dependencies extend across years as well as months.

Let's start with the differencing parameter. There appears to be at least some differencing needed because the series drifts across time (overall level of passengers increases). The ACF and PACF plots give us more detailed information on this.

- Select **Series/ACF plot...** from the **Stats** menu
- Select the variable to be plotted: PASS
- Click OK

```
Plot of     PASS
Number of cases =  144
Mean of series =       5.542
Standard deviation of series =       0.440

Plot of Autocorrelations

 Lag   Corr     SE -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
                     |____|____|____|____|____|____|____|____|____|____|
   1   .954   .083                          (    |...}.................
   2   .899   .140                        (      |.....}...............
   3   .851   .175                      (        |.......}.............
   4   .808   .202                    (          |.........}...........
   5   .779   .223                   (           |..........}..........
   6   .756   .242                  (            |...........}.........
   7   .738   .257                  (            |...........}.........
   8   .727   .272                 (             |............}........
   9   .734   .285                (              |.............}.......
  10   .744   .298                (              |.............}.......
  11   .758   .310               (               |..............}......
  12   .762   .323              (                |...............}.....
  13   .717   .335              (                |...............}.....
  14   .663   .346             (                 |................}....
  15   .618   .354             (                 |................ }...
```


© 1989, SYSTAT, Inc.

- Select **Series/PACF plot...** from the **Stats** menu
- Select the variable to be plotted: PASS
- Click OK

```
Plot of     PASS
Number of cases =  144
Mean of series =       5.542
Standard deviation of series =       0.440

Plot of Partial Autocorrelations

 Lag   Corr     SE -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
                     |____|____|____|____|____|____|____|____|____|____|
   1   .954   .083                          (    |...}.................
   2  -.118   .083                          (  ..|    )
   3   .054   .083                          (    |.   )
   4   .024   .083                          (    |    )
   5   .116   .083                          (    |..  )
   6   .044   .083                          (    |.   )
   7   .038   .083                          (    |    )
   8   .100   .083                          (    |..  )
   9   .204   .083                          (    |...}.
  10   .064   .083                          (    |.   )
  11   .106   .083                          (    |..  )
  12  -.042   .083                          (  .-|    )
  13  -.485   .083                    ......(...|    )
  14  -.034   .083                          (    |    )
  15   .042   .083                          (    |.   )
```

Notice that the autocorrelations are substantial and well outside two standard errors on the plot. There are two bulges in the ACF plot at Lag=1 and Lag=12, suggesting the nonseasonal (monthly) and seasonal (yearly) dependencies we supposed. The PACF plot shows the same dependencies more distinctly. Let's difference the series once to see if the overall trend (drift) goes away. We'll do the same two plots.

- Select **Series/Transform...** from the **Stats** menu
- Choose the type of transformation: [Difference]
- Select the variable: PASS
- Click OK

© 1989, SYSTAT, Inc.

- Select **Series/ACF plot...** from the **Stats** menu
- Select the variable to be plotted: PASS
- Click OK

```
Plot of     PASS
Number of cases =  143
Mean of series =       0.009
Standard deviation of series =       0.106

Plot of Autocorrelations

 Lag   Corr     SE -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
                      |____|____|____|____|____|____|____|____|____|____|
   1   .200   .084                        (    |***)
   2  -.120   .087                       (***  |   )
   3  -.151   .088                       (***  |   )
   4  -.322   .090                  ****(***  |   )
   5  -.084   .098                       ( **  |   )
   6   .026   .098                       (     |   )
   7  -.111   .098                       ( **  |   )
   8  -.337   .099                  ****(***  |   )
   9  -.116   .107                       (  **  |   )
  10  -.109   .108                       (  **  |   )
  11   .206   .108                       (      |****)
  12   .841   .111                       (      |****)**************
  13   .215   .149                  (           |***** )
  14  -.140   .151                  (    ***    |      )
  15  -.116   .152                  (     **    |      )
```

- Select **Series/PACF plot...** from the **Stats** menu
- Select the variable to be plotted: PASS
- Click OK

```
Plot of     PASS
Number of cases =  143
Mean of series =       0.009
Standard deviation of series =       0.106

Plot of Partial Autocorrelations

 Lag   Corr     SE -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
                      |____|____|____|____|____|____|____|____|____|____|
   1   .200   .084                        (    |***)
   2  -.167   .084                       (***  |   )
   3  -.096   .084                       ( **  |   )
   4  -.311   .084                  ***(***  |   )
   5   .008   .084                       (     |   )
   6  -.075   .084                       (  *  |   )
   7  -.210   .084                      .(***  |   )
   8  -.495   .084                ********(***  |   )
   9  -.192   .084                       (***  |   )
  10  -.532   .084                *********(***  |   )
  11  -.302   .084                  ***(***  |   )
  12   .586   .084                       (    |***)**********
  13   .026   .084                       (     |   )
  14  -.181   .084                       (***  |   )
  15   .120   .084                       (    |***)
```


© 1989, SYSTAT, Inc.

Now we have only 143 points in the series because the first point had no prior value to remove. It was therefore set to missing. The two plots show that the differencing has substantially removed the monthly changes in trend. We still have the seasonal (yearly) trend, however. We therefore difference again and replot:

- Select **Series/Transform...** from the **Stats** menu
- Choose [Difference]
- Specify Lag=12
- Select the variable: PASS
- Click OK

- Select **Series/ACF plot...** from the **Stats** menu
- Select the variable to be plotted: PASS
- Click OK

```
Plot of     PASS
Number of cases =  131
Mean of series =       0.000
Standard deviation of series =       0.046

Plot of Autocorrelations

 Lag   Corr     SE -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
                      |____|____|____|____|____|____|____|____|____|____|
   1  -.341   .087                  ****(***  |   )
   2   .105   .097                       (     |** )
   3  -.202   .098                      .(***  |   )
   4   .021   .101                       (     |    )
   5   .056   .101                       (     |*   )
   6   .031   .101                       (     |    )
   7  -.056   .101                       (     *    )
   8  -.001   .102                       (     |    )
   9   .176   .102                       (     |****)
  10  -.076   .104                       (     *    )
  11   .064   .104                       (     |*   )
  12  -.387   .105                  ****(****  |    )
  13   .152   .115                       (     |***  )
  14  -.058   .117                       (     *    )
  15   .150   .117                       (     |***  )
```

© 1989, SYSTAT, Inc.

- Select **Series/PACF plot...** from the **Stats** menu
- Select the variable to be plotted: PASS
- Click OK

```
Plot of      PASS
Number of cases =  131
Mean of series =        0.000
Standard deviation of series =        0.046

Plot of Partial Autocorrelations

 Lag   Corr     SE -1.0  -.8  -.6  -.4  -.2   .0   .2   .4   .6   .8  1.0
                      |----|----|----|----|----|----|----|----|----|----|
   1  -.341   .087                      ****{***|    )
   2  -.013   .087                           {   |    )
   3  -.193   .087                           {***|    )
   4  -.125   .087                           {***|    )
   5   .033   .087                           {   |    )
   6   .035   .087                           {   |    )
   7  -.060   .087                           {  .|    )
   8  -.020   .087                           {   |    )
   9   .226   .087                           {   |***).
  10   .043   .087                           {   |*   )
  11   .047   .087                           {   |*   )
  12  -.339   .087                      ****{***|    )
  13  -.109   .087                           { **|    )
  14  -.077   .087                           {  *|    )
  15  -.022   .087                           {   |    )
```

Most of the dependency seems to have been removed. Although there are some autocorrelations and partial autocorrelations outside two standard errors, we will not difference again. We will fit a model first because over-differencing can mask the effects of MA parameters. In fact, the pattern in this last plot suggests one regular and one seasonal MA parameter because there are ACF spikes (instead of bulges) at lags 1 and 12, and the PACF shows decay at lags 1 and 12. You should consult the references previously cited for more information on how to read these plots for identification.

9.4 ARIMA with backcasting

Here we fit a seasonal multiplicative ARIMA model with no autoregressive parameter, one difference parameter, one moving average, no seasonal autoregressive parameter, one seasonal difference parameter, and one seasonal moving average parameter. Recall that we have already differenced the series twice. Therefore, both difference parameters are done.

© 1989, SYSTAT, Inc.

- Select **Series/ARIMA...** from the **Stats** menu
- Select the variable to be analyzed: PASS
- Specify values for the parameters: 1 for [MA parameters] and 1 for [Seasonal MA] parameters
- Specify Seasonal periodicity 12
- Specify number of cases to be [Backcast]: 13
- Click OK

The [Season] option tells SYSTAT that the seasonal parameters have a seasonal period of 12 months. The [Backcast] option helps us compute a better estimate of the MA parameters. You will have to read more about it in the references, but it extends the series backwards (forecasting in reverse) so as to approximate a continuous function better. Although it slows down computation, you should use backcasting for seasonal models especially and choose a length greater than the seasonal period. We didn't use the [Constant] option for the airline data because the mean of the differenced logged series was near zero. This is often true for differenced series.

```
Iteration  Sum of Squares  Parameter Values
    0        .2392764D+00   .100  .100
    1        .1835532D+00   .345  .433
    2        .1764962D+00   .449  .633
    3        .1759952D+00   .416  .592
    4        .1758742D+00   .409  .613
    5        .1758463D+00   .392  .614
    6        .1758443D+00   .396  .613
    7        .1758443D+00   .396  .613
    8        .1758443D+00   .396  .613
    9        .1758443D+00   .396  .613
   10        .1758443D+00   .396  .613

Final Value OF MSE is        0.001

Index   Type    Estimate       A.S.E.       Lower  <95%> Upper

  1       MA       0.396        0.093        0.212        0.579
  2      SMA       0.613        0.074        0.467        0.760

Asymptotic Correlation Matrix of Parameters
                      1            2

          1         1.000
          2        -0.171        1.000
```

After the ARIMA command has been executed, the residuals are stored in the active work area under the variable name RESIDUAL. Since it is not a variable in your file, it does not appear in selection lists in any of the dialog boxes. You can access it with **Series** and other **Stats** and **Graph** menu items (for plotting or further analysis) by typing RESIDUAL where you want it.

© 1989, SYSTAT, Inc.

The original series is not altered. This means that you can check the adequacy of the model (the third Diagnosis stage of ARIMA modeling) by using the various facilities of the **Stats** menu. In addition, if you use the [Save smoothed values] option, the residuals are saved in a SYSTAT file. You can then do normal probability plots, stem-and-leaf plots, Kolmogorov-Smirnov tests, and other statistical tests on residuals.

Let's focus on the question of serial dependence among the residuals.

- Select **Series/ACF plot...** from the **Stats** menu
- Select the variable to be plotted: RESIDUAL
- Click OK

```
Plot of RESIDUAL
Number of cases =  131
Mean of series =        0.000
Standard deviation of series =        0.036

Plot of Autocorrelations
```

Lag	Corr	SE	-1.0 -.8 -.6 -.4 -.2 .0 .2 .4 .6 .8 1.0
1	.017	.087	(|)
2	.019	.087	(|)
3	-.126	.087	(*** |)
4	-.142	.089	(*** |)
5	.050	.091	(| *)
6	.062	.091	(| *)
7	-.073	.091	(* |)
8	-.038	.091	(|)
9	.103	.092	(| **)
10	-.079	.092	(* |)
11	.025	.093	(|)
12	-.010	.093	(|)
13	.032	.093	(|)
14	.043	.093	(| *)
15	.048	.093	(| *)

The autocorrelation plot (ACF) is quite well behaved.

© 1989, SYSTAT, Inc.

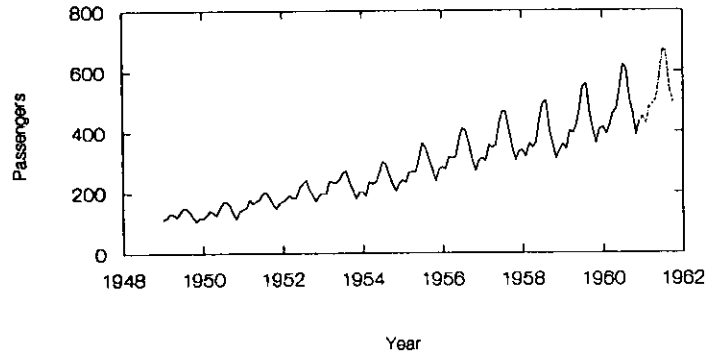
9.5 ARIMA with forecasting

The [Forecast] option of the ARIMA command can forecast future values in a series (or forecast values in the series itself for diagnosing the model). We could have added forecasting by specifying 10 cases to be forecast. We would have gotten the following additional information:

Forecast Values

PERIOD	LOWER95	FORECAST	UPPER95
1961.01	418.862	450.296	484.090
1961.02	372.894	405.792	441.592
1961.03	376.411	413.998	455.339
1961.04	429.413	476.830	529.482
1961.05	424.387	475.395	532.532
1961.06	487.259	550.278	621.447
1961.07	554.689	631.220	718.309
1961.08	524.220	600.854	688.691
1961.09	442.105	510.209	588.805
1961.10	403.058	468.191	543.849

The "forecast origin" in this case is taken as the last point in the series. From there, the model computes and prints 10 new points with their upper and lower 95 percent asymptotic confidence intervals. You can save the forecasts by clicking [Save forecasts] and then plot them with **Graph**. Here is a plot of the original series with the forecast cases shown as dotted lines.



Incidentally, the SYSTAT decision to put RESIDUAL into your active work area automatically after ARIMA was based on the need to always plot residuals after an analysis. Since plotting forecasts can wait until after you have correctly identified a model, it was decided to leave the active work area free for further modeling steps.

© 1989, SYSTAT, Inc.

9.6 Exponential smoothing

In the previous example, we looked at the AIRLINE data from Box and Jenkins. The plot of the series shows a strong increasing trend and what looks like multiplicative seasonality. We could try to forecast this series with a model having a linear trend and multiplicative seasonality:

- Select **Series/Exponential...** from the **Stats** menu
- Choose the components and weights for the smoothing: Location weight.3, [Linear] trend component with weight .4, and [Multiplicative] season component with weight .4
- Specify number of cases to be forecast: 10
- Select the variable to be smoothed: PASS
- Click OK

The output begins with the model and initial parameter estimates:

```
Smooth Location Parameter with Coefficient= .300
Linear Trend with Smoothing Coefficient= .400
Multiplicative Seasonality with Smoothing Coefficient= .400

Initial values

Seasonal Indices for the Series are:
   1:      91.077
   2:      88.133
   3:     100.825
   4:      97.321
   5:      98.305
   6:     111.296
   7:     122.636
   8:     121.652
   9:     105.997
  10:      92.200
  11:      80.397
  12:      90.164


Initial Smoothed Value  =      88.263
Initial Trend Parameter =       2.645
```

The initial seasonal indices are the same as those obtained using **Seasonal decomposition...** and are interpreted and used the same way. The initial smoothed value is a regression estimate of the level of the seasonally adjusted series immediately before the first observation in the sample. The initial trend parameter is the slope of the regression of observations on observation number—the increase or decrease from one observation to the next due to the overall trend. For a percentage growth model the trend parameter is the expected percentage change from the previous to the current observation due to trend.

© 1989, SYSTAT, Inc.

After the values are smoothed, SYSTAT prints the final estimates of the seasonal, location, and trend parameters, plus the within-series forecast error:

```
Final Values

Seasonal Indices for the Series are:

   1:      87.628
   2:      82.996
   3:      95.362
   4:      98.155
   5:     101.936
   6:     117.471
   7:     133.389
   8:     130.644
   9:     107.631
  10:      93.352
  11:      78.956
  12:      86.086

Final Smoothed Value  =     497.921
Final Trend Parameter =       8.374

Within Series MSE=     325.915 SE=      18.053
```

You can vary the smoothing coefficients and see if they reduce the standard error. For example, if we used equal weights .2, .2 and .2, we would get a smaller within-series forecast error:

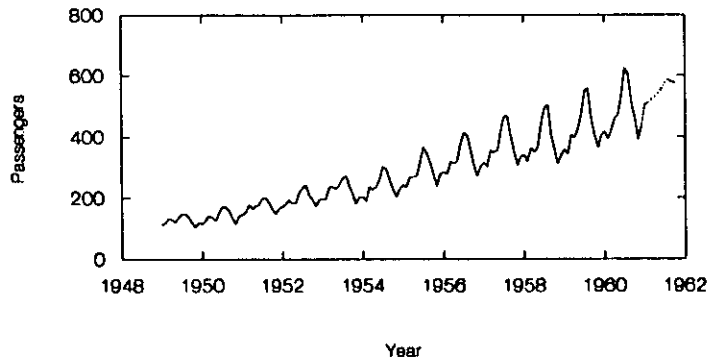
```
Within Series MSE=     220.026 SE=      14.833
```

Finally we get the out-of-series forecasts:

```
    OBS       FORECAST

1961.01         505.260
1961.02         511.822
1961.03         521.879
1961.04         530.800
1961.05         540.603
1961.06         556.945
1961.07         576.114
1961.08         585.445
1961.09         579.041
1961.10         576.097
```

© 1989, SYSTAT, Inc.

Here is a plot of the forecasts:



In-series forecasts

Sometimes it's best to develop a model on a portion of a series and see how well it predicts the remainder. There are twelve years of airline data for a total of 144 monthly observations. The following command develops the smoothing model with the first ten years of data (120 observations) and predicts the final two years (observations 121–144):

- Select **Series/Exponential...** from the **Stats** menu
- Choose the components and weights for the smoothing: [location with weight .3, [Linear] trend component with weight .4, and [Multiplicative] season component with weight .4
- Specify range of cases to be forecast: 121 to 144
- Select the variable to be smoothed: PASS
- Click OK

© 1989, SYSTAT, Inc.

Output from this procedure includes the following forecasts.

```
Within Series MSE=     292.090 SE=      17.091

   OBS       FORECAST
 1959.02        415.309
 1959.03        422.605
 1959.04        428.775
 1959.05        435.905
 1959.06        447.866
 1959.07        459.765
 1959.08        466.434
 1959.09        463.163
 1959.10        461.209
 1959.11        458.712
 1959.12        468.994
 1960.01        476.594
 1960.02        479.834
 1960.03        498.645
 1960.04        504.144
 1960.05        513.823
 1960.06        538.906
 1960.07        559.431
 1960.08        563.293
 1960.09        543.009
 1960.10        529.579
 1960.11        517.247
 1960.12        533.425

Forecast MSE=    4341.683 SE=      65.891
```

Note that the within-series standard error is not the same as in the previous run because it's now based on only the first 120 observations. The error for the actual forecasts (65.891) is much larger than that for the in-series forecasts (17.091).

For a thorough review of issues and developments in exponential smoothing models, see Gardner (1985). For an introduction to these models, see any introductory forecasting book such as Makridakis, Wheelwright, and McGee (1983).

9.7 Seasonal decomposition and adjustment

We use here the same airline data from Box and Jenkins (1976) used in Example 9.3. If you examine the plot there, you can see the strong periodicities. The size of the periodicities depends on the level of the series, so we know that the form of seasonality is multiplicative. Notice also that these cycles do not seem to be smooth functions of sine and cosine waves like the Newark temperature data in Example 9.8. Instead, there are jagged spikes in the data that correspond, apparently, to holidays like Christmas and Easter.

© 1989, SYSTAT, Inc.

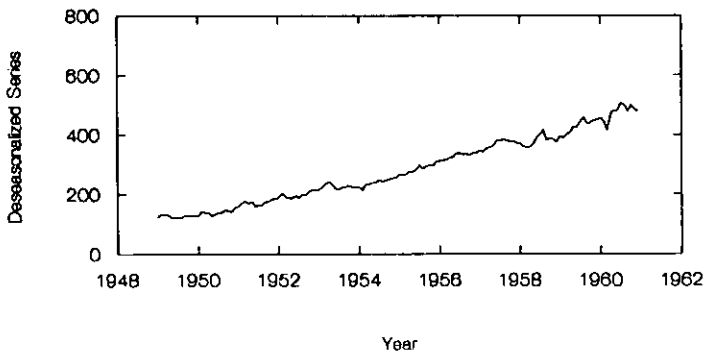
Adjust the airline series for the multiplicative seasonal effect implied by the series plot:

- Select **Series/Seasonal...** from the **Stats** menu
- Choose [Multiplicative] effects
- Select the variable to be smoothed: PASS
- Click OK

```
Adjust Series for a Seasonal Periodicity of   12

Seasonal Indices for the Series are:

   1:      91.077
   2:      88.133
   3:     100.825
   4:      97.321
   5:      98.305
   6:     111.296
   7:     122.636
   8:     121.652
   9:     105.997
  10:      92.200
  11:      80.397
  12:      90.164
Series is Transformed
```

Airline travel appears heaviest during the summer months, June (6) through September (9). We can now plot the adjusted series.

- Select **Series/Case plot...** from the **Stats** menu
- Specify the variable to be plotted: PASS
- Click OK

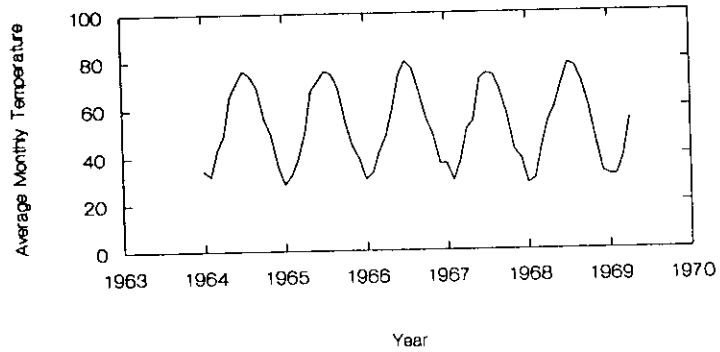


The plot shows that the trend and the irregular components remain, but the seasonal component has been removed from the series.

© 1989, SYSTAT, Inc.

9.8 Fourier analysis

Let's look at a typical Fourier application. The data in the NEWARK file are 64 average monthly temperatures in Newark, New Jersey beginning January, 1964. The data are from the U.S. Government, cited in Chambers *et al.* (1983). Notice that their fluctuations look something like a sine wave, so we might expect that they could be modeled adequately by the sum of a relatively small number of trigonometric components. We have taken exactly 64 measurements to fulfill the powers of 2 rule.



© 1989, SYSTAT, Inc.

- Select **Series/Transform...** from the **Stats** menu
- Choose the [Mean] transformation
- Select the variable to be analyzed: TEMP
- Click OK

- Select **Fourier...** from the **Stats** menu
- Select the variable to be transformed: TEMP
- Click OK

Fourier Components of TEMP

Index	Frequency	Real	Imaginary	Magnitude	Phase	Periodogram
1	0.00000	0.000	0.000	0.000	0.000	0.000
2	0.01563	-0.763	-0.363	0.845	-2.697	11.416
3	0.03125	-0.803	-0.177	0.822	-2.924	10.807
4	0.04688	-1.587	-0.779	1.768	-2.685	50.016
5	0.06250	-1.658	-1.817	2.460	-2.310	96.810
6	0.07813	-6.248	-7.214	9.544	-2.285	1457.410
7	0.09375	2.606	3.633	4.471	0.948	319.813
8	0.10938	1.040	1.786	2.067	1.044	68.359
9	0.12500	0.592	0.936	1.107	1.007	19.618
10	0.14063	0.438	0.588	0.733	0.930	8.604
11	0.15625	-0.127	1.135	1.142	1.682	20.859
12	0.17188	0.067	0.715	0.718	1.477	8.252
13	0.18750	-0.255	0.785	0.825	1.885	10.885
14	0.20313	0.140	0.132	0.192	0.756	0.588
15	0.21875	-0.071	0.291	0.299	1.811	1.432

An easier way to process this information is to plot magnitude with the **Case plot...** We can do this because the index number of the series now indexes frequency rather than time. A periodogram is a plot of the square of magnitude against frequency. Remember that magnitude is stored where TEMP used to be. Don't worry. We can get the original temperatures back by opening the file again or using **Clear**. Anyway, here's how to get the first 15 components (if you want more, use the [Lag] option):

- Select **Series/Transform...** from the **Stats** menu
- Choose the transformation to be performed: [Square]
- Select the variable to be transformed: TEMP
- Click OK

- Select **Series/Case plot...** from the **Stats** menu
- Select the variable to be plotted: TEMP
- Click OK

© 1989, SYSTAT, Inc.


```
Series is transformed

Plot of      TEMP
Number of Cases =   64
Mean of Series =       4.169
Standard Deviation of Series =       16.019

Sequence Plot of Series

 Case        Value          0.000                                                91.08
                            +----+----+----+----+----+----+----+----+----+----+----+
        1        0.000       *
        2        0.713       *
        3        0.675       *
        4        3.126       **
        5        6.051       ****
        6       91.088       **************************************************
        7       19.988       ***********
        8        4.272       ***
        9        1.226       *
       10        0.538       *
       11        1.304       *
       12        0.516       *
       13        0.680       *
       14        0.037       *
       15        0.090       *
```

Notice, our hunch was largely correct. There is one primary peak at a relatively low frequency. Two final points. First, some analysts prefer to plot the logs of these values against frequency. We could do this in the following way:

- Select **Series/Transform...** from the **Stats** menu
- Choose the transformation to be performed: [Square]
- Select the variable to be transformed: TEMP
- Click OK

- Select **Series/Transform...** from the **Stats** menu
- Choose the transformation to be performed: [Log]
- Click OK

- Select **Series/Case plot...** from the **Stats** menu
- Select the variable to be plotted: TEMP
- Click OK

Logging, by the way, looks noisier than the plot above, but can reveal significant spikes that might be hidden in the raw periodogram.

The second point involves smoothing the periodogram. Often it is best to taper the series first before computing the periodogram. This makes the "spikes" more pronounced in the log-periodogram plot.

© 1989, SYSTAT, Inc.

Examples

- Select **Series/Transform...** from the **Stats** menu
- Choose the transformation to be performed: [Mean]
- Click OK

- Select **Series/Transform...** from the **Stats** menu
- Choose the transformation to be performed: [Split-cosine-bell]
- Click OK

- Select **Series/Fourier...** from the **Stats** menu
- Select the variable to be transformed: TEMP
- Click OK

- Select **Series/Transform...** from the **Stats** menu
- Choose the transformation to be performed: [Square]
- Click OK

- Select **Series/Transform...** from the **Stats** menu
- Choose the transformation to be performed: [Log]
- Click OK

- Select **Series/Case plot...** from the **Stats** menu
- Click OK

© 1989, SYSTAT, Inc.

Examples

Series

© 1989, SYSTAT, Inc.

10 Stats

Introduction 446
Location 446
Spread 447
The normal distribution 447
Nonnormal shape 448
Subpopulations 450
One-way ANOVA and t-tests 451
Usage 453
Statistics... 453
Statistics by group 454
Saving statistics 454
t-test... 455
Independent *t*-tests 456
Paired *t*-tests 456
Matched pairs 456
One-sample *t*-tests 457
One-way ANOVA... 457
Post hoc pairwise comparisons 458
Computation 459
Missing data 459
Algorithms 459
Examples 460
10.1 Basic statistics 460
10.2 Statistics by groups 460
10.3 Two sample *t*-test 462
10.4 Dependent *t*-test 463
10.5 Analysis of variance 464

© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

Overview

Stats does the following:

1) It computes a wide variety of descriptive statistics. If you specify a grouping variable or variables, you can get statistics within subgroups or cells of your data. Any statistics calculated can be saved into a SYSTAT file so you can aggregate data within groups.

2) It performs *t*-tests, including two-sample (independent), paired (dependent), and one-sample tests.

3) It produces a one-way analysis of variance with Tukey HSD post hoc tests of pairwise mean differences.

Descriptive statistics are numerical summaries of batches of numbers. Inevitably, these summaries are misleading, because they mask details of the data. Without them, however, we would be lost in particulars.

There are many ways to describe a batch of data. Not all are appropriate for every batch, however. Let's look at the *Who's Who* data from the SYSTAT general introduction to see what this means. First of all, here is a stem and leaf diagram of the ages of 50 randomly sampled people from *Who's Who*. Remember, a stem and leaf diagram is a tally. It shows us the distribution of the AGE values.

```
        Stem and leaf plot of variable:      AGE     , N =    50

Minimum:              34.000
Lower hinge:          49.000
Median:               56.000
Upper hinge:          66.000
Maximum:              81.000

                3   4
                3   689
                4   14
                4 H 556778999
                5   0011112
                5 M 556688889
                6   0023
                6 H 55677789
                7   04
                7   5668
                8   1
```

Notice that these data look fairly symmetric and lumpy in the middle. A natural way to describe this type of distribution would be to report its center and the amount of spread.

Location

How do we describe the center, or central location of the distribution on a scale? One way is to pick the value above which half the data values fall and, by implication, below which half the data values fall. This measure is called the *median*. For our AGE data, the median age is 56 years. Another measure of location is the "center of gravity" of the numbers. Think of turning the stem and leaf diagram on its side and balancing it. The balance point would be the *mean*. For a batch of numbers the mean is computed by averaging the values. In our sample, the mean age is 56.7 years. It is quite close to the median.

© 1989, SYSTAT, Inc.

Spread

One way to measure spread would be to take the difference between the largest and smallest value in the data. This is called the *range*. For the age data, the range is 47 years. Another measure, called the *interquartile range* or *midrange*, is the difference between the values at the limits of the middle 50 percent of the data. For AGE, this is 17 years. Still another way to measure would be to compute the average variability in the values. The *standard deviation* is the square root of the average squared deviation of values from the mean. For the AGE variable, the standard deviation is 11.62. The stem and leaf diagram in SYGRAPH and that produced by **Stats** display these statistics. Here is some output from **Stats** for our data:

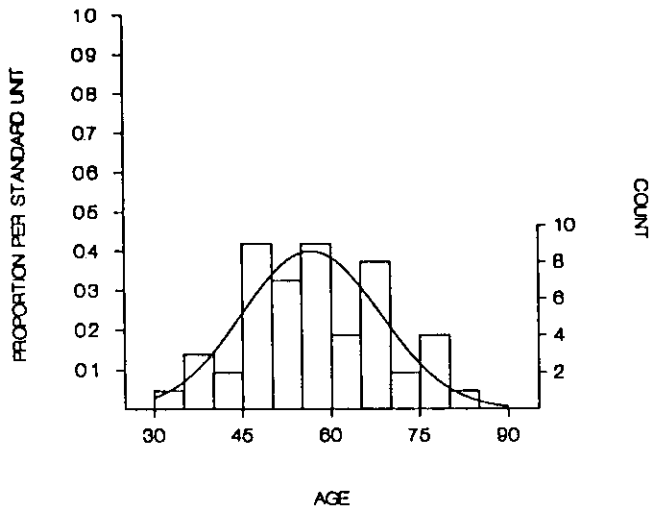
```
                    AGE

N of cases           50
Mean              56.700
Standard Dev      11.620
```

The normal distribution

All these measures of location and spread have their advantages and disadvantages, but the mean and standard deviation are especially useful for describing data that follow a *normal distribution*. The normal distribution is a mathematical curve with only two parameters in its equation: the mean and standard deviation. As you recall from the **Introduction to statistics** chapter, a parameter defines a family of mathematical functions, all of which have the same general shape. Thus, if data come from a normal distribution, we can describe them completely (except for random variation) with only a mean and standard deviation.

Let's see how this works for our AGE data. Shown in the next figure is a histogram of AGE with the normal curve superimposed. The location (center) of this curve is at the mean age of the sample (56.7) and its spread is determined by the standard deviation (11.62).

© 1989, SYSTAT, Inc.

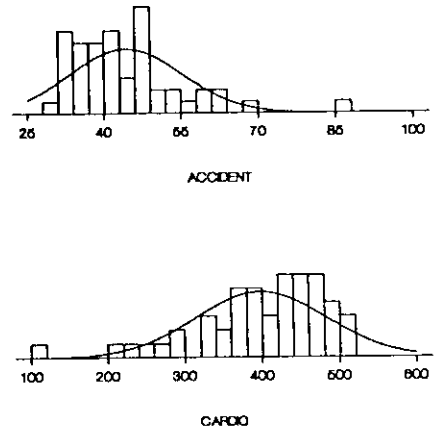


The fit of the curve to the data looks excellent. Let's examine the fit in more detail For a normal distribution, we would expect 68 percent of the observations to fall between one standard deviation below the mean and one standard deviation above the mean (45.1 to 68.3 years). By counting values in the stem and leaf diagram, we find 34 cases—on target. This is not to say that every number follows a normal distribution exactly, however. If we looked further, we would find that the tails of this distribution are slightly shorter than those from a normal distribution, but not enough to worry.

Nonnormal shape

Before you compute means and standard deviations on everything in sight, however, let's take a look at some more data: the MEDICAL data used in Example 10.1. Here are histograms for the first two variables, ACCIDENT and CARDIO:

© 1989, SYSTAT, Inc.



Notice that the normal curves fit the distributions poorly. ACCIDENT is positively skewed. That is, it has a long right tail. CARDIO, on the other hand, is negatively skewed. It has a long left tail. The means (44.3 and 398.5) calculated in Example 10.1 clearly do not fall in the centers of the distributions. Furthermore, if you calculate the medians using the **Stem** display, you will see that the mean for ACCIDENT is pulled away from the median (41.9) toward the upper tail and the mean for CARDIO is pulled to the left of the median (416.2).

In short, means and standard deviations are not good descriptors for nonnormal data. In these cases, you have two alternatives: either transform your data to look normal, or find other descriptive statistics that characterize the data. If you log the values of ACCIDENT, for example, the histogram looks quite normal. If you square the values of CARDIO, the normal fit similarly improves.

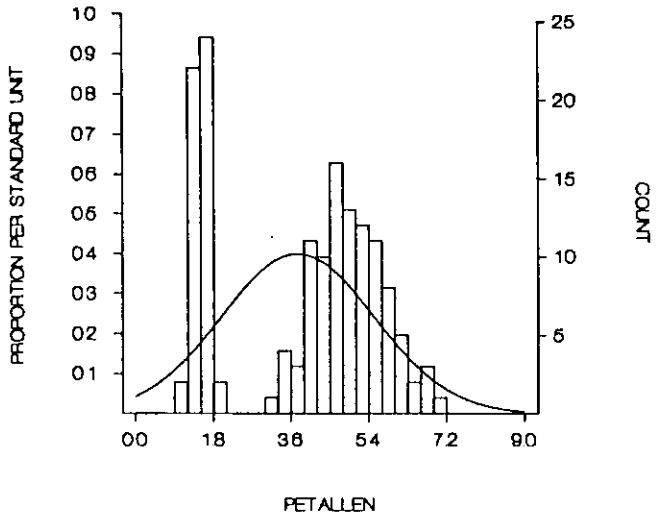
© 1989, SYSTAT, Inc.

If a transformation doesn't work, then you may be looking at data that come from a different mathematical distribution or are mixtures of subpopulations (see below). The probability plots in SYSTAT can help you identify certain mathematical distributions. There is not room here to discuss parameters for more complex probability distributions.

Otherwise, you should turn to distribution-free summary statistics to characterize your data: the median, range, minimum, maximum, midrange, quartiles, and percentiles. You can calculate these statistics with **Graph/Stem**, with **Stats/Statistics...**, and with the [List format] option for **Tables/Tabulate...**

Subpopulations

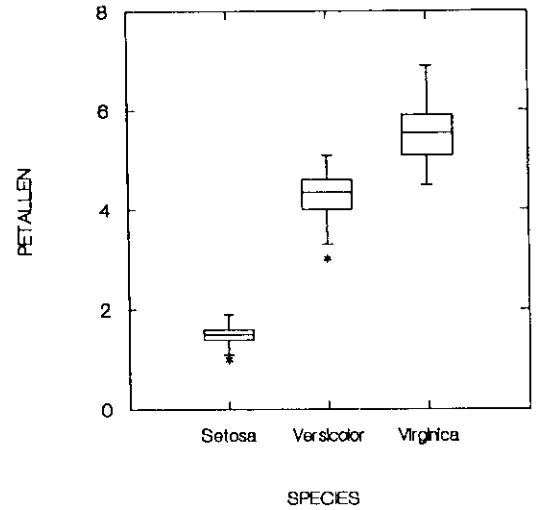
Sometimes distributions can look nonnormal because they are mixtures of different normal distributions. Let's look at the Fisher/Anderson IRIS flower measurements from Example 10.5. Here is the histogram of PETALLEN (petal length) smoothed by a normal curve:



We forgot to notice that the petal length measurements involve three different flower species. You can see one of them at the left. The other two are blended at the right. Computing a mean and standard deviation on the mixed data is misleading.

© 1989, SYSTAT, Inc.

A box plot split by species shows how different the subpopulations are.



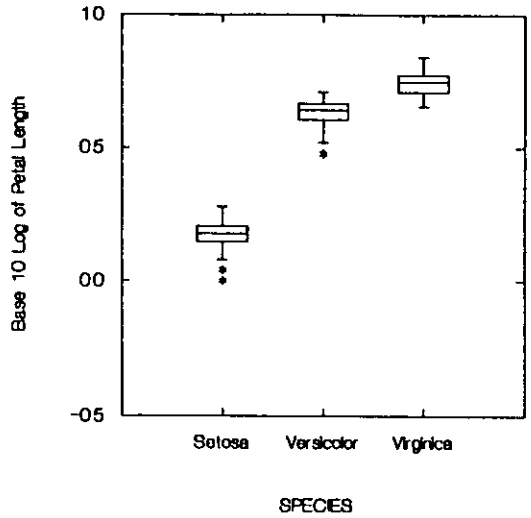
When there are such differences, you should compute basic statistics by group. You can do this with **By groups...** from the **Data** menu.

If you wish to go on to test whether the differences in subpopulation means are significant, the **One-way ANOVA** option in **Stats** automatically computes the summary statistics by group. See Example 10.5.

One-way ANOVA and *t*-tests

These procedures are designed to test the significance of the difference between two means (*t*-tests) or among several means (one-way ANOVA). Both procedures assume that the data are normally distributed within groups and that each group has the same standard deviation You can see the examples later in this chapter for more information. Notice, however, that the spreads differ for the three flower species on PETALLEN. This makes *t*-test between any two species on this variable somewhat questionable and an ANOVA across the groups similarly shaky. When the standard deviations are proportional to the cell means, as in this case, a log transformation of the data may remedy the problem. Here is a box plot of the same data after logging to the base 10 with the transformation LOG(PETALLEN)/LOG(10).

© 1989, SYSTAT, Inc.

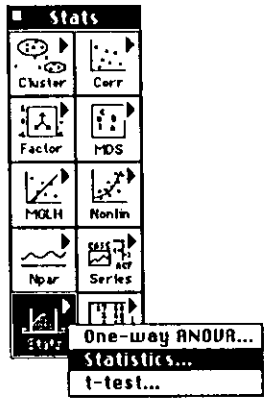


Repeat the ANOVA using this transformation on the IRIS data and see how the results differ from those calculated in Example 10.5.

© 1989, SYSTAT, Inc.

Usage

There are three submenus: **Stats/Statistics...** for descriptive statistics, **Stats/t-test...** for *t*-tests, and **Stats/One-way ANOVA** for one-way analysis of variance.



Press and hold the mouse button on **Stats**. A submenu pops up offering **One-way ANOVA...**, **Statistics...**, and **t-test...** Slide the pointer to the item you want and release the mouse button.

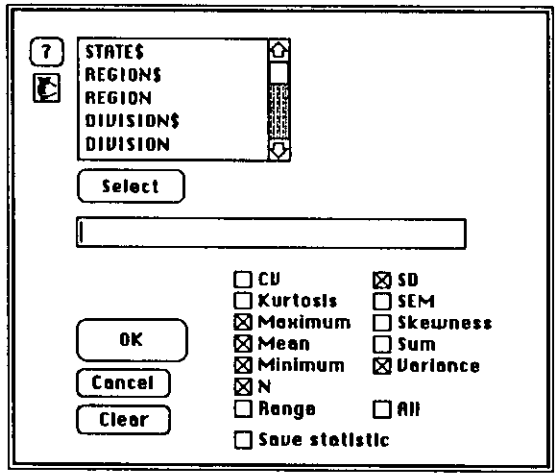
Statistics...

For each numeric variable, **Stats** computes the mean, standard deviation, minimum, maximum, range, sum, standard error of the mean, skewness, kurtosis, and coefficient of variation. The default output is number of cases, minimum, maximum, mean, and standard deviation.

The standard deviation measures the spread of a batch, skewness measures its symmetry, and kurtosis its peakedness. The measures of skewness and kurtosis, g1 and g2,are zero centered. A positive value for skewness indicates a long right tail and a negative value, a long left tail. A positive value for kurtosis indicates a longer tailed than normal distribution and a negative value, a shorter tailed. The standard error of g1 is $\sqrt{6/N}$ and the standard error of g2 is $\sqrt{24/N}$. CV is the coefficient of variation (SD/Mean), and SEM is the standard error of the mean ($SD/\sqrt{N}$).

© 1989, SYSTAT, Inc.

- Select **Stats/Statistics...** from the **Stats** menu
  - Choose the statistics to be computed; [Maximum], [Mean], [Minimum], [N], [SD], and [Variance] are the defaults
  - Select the variables to be analyzed; all numeric variables are used by default
- Click OK



Statistics by groups

You can do basic statistics by groups with **Data/By groups...** First, you must sort the data by the grouping variable(s) with **Sort** from the **Data** menu before using **Statistics...**. This sorting requirement may seem inconvenient, but it keeps your computer's finite memory from limiting the number of groups and counts within groups.

Saving statistics

If you have a file sorted by one or more grouping variables and want to produce another file containing aggregate statistics, click [Save statistic]. Whatever statistic you choose is saved into the new file. There is one case in the new file for each group. This way, you can save aggregate means, standard deviations, or any other statistic. You may save only one type of statistic, however, in a single file (the first or only option you choose in the dialog box). If you want to analyze more than one type of aggregate statistic from the same data, use Stats again and save another file.

© 1989, SYSTAT, Inc.

Here is an example for the MEDICAL data.

- Select **Stats/Statistics...** from the **Stats** menu
- Choose the statistic to be computed; [Mean]
- Select a grouping variable: DIVISION
- Click [Save statistic] to save means in a file named MEANMED
- Click OK

MEANMED contains 9 cases (from the 9 census regions), each being the means on every numeric variable for the appropriate region.

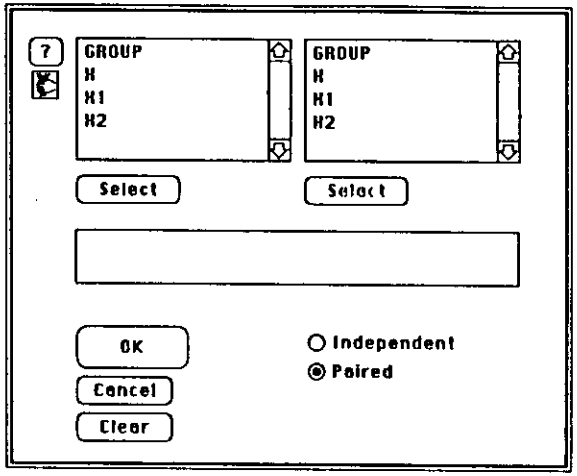
Aggregate statistics can be merged back into your original file by using the grouping variable you selected. For example, Levene's test for the homogeneity of variance requires a one-way analysis of variance on the absolute deviations from cell means (Levene, 1960).

t-test...

SYSTAT can compute independent or dependent *t*-tests. For independent tests, you must have a grouping variable with exactly two values. The variable may be either numeric or character. The file need not be sorted on the grouping variable. For dependent tests, SYSTAT tests every pairing of the variables you select. The dependent *t*-test uses pairwise deletion of missing values when you select more than two variables.

- Select **Stats/t-test...** from the **Stats** menu
  - Choose the type of test to be performed
- Select the variables to be analyzed: for [Independent], choose one or more variables; for [Paired], choose two or more variables
  - For [Independent], select a grouping variable
- Click OK

© 1989, SYSTAT, Inc.



Independent *t*-tests

For independent or two-sample tests, you must select one or more variables to test from the left selection list and one grouping variable from the right selection list. Then click the [Independent] button. Stats prints group means, standard deviations, and sample sizes. It tests equality of group means with and without the assumption of equal variances (the former is the pooled variances *t*-test, the latter the separate variances test).

Paired *t*-tests

For paired tests, click the [Paired] button and select variables from the left selection list only. Stats reports the average difference in means, the standard deviation of the differences, and the *t* statistic along with its associated *p* value.

Matched pairs

The mechanics of this test are identical to those for the paired test. The only difference is in your study design. The matched design pairs different subjects with common values on one or more matching variables, while the paired test uses two measures on the same subject as a pair. In a matched case-control design, for example, make sure you have the case and control measurements on the same record in your file.

© 1989, SYSTAT, Inc.

One-sample *t*-tests

You can do a one-sample test by adding a column to your data file that contains a constant value corresponding to the population mean of your null hypothesis. Then perform a paired *t*-test on this constant and the variable you are testing.

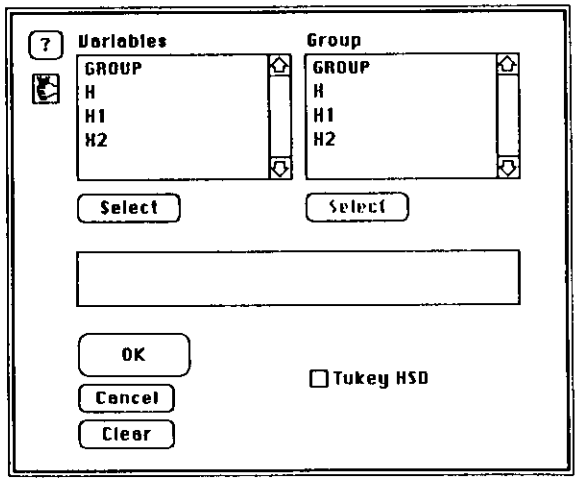
One-way ANOVA...

One-way analysis of variance is a generalization of the two-sample *t*-test to more than two groups. For a one-way ANOVA, Stats reports descriptive statistics and an ANOVA. You may use a numeric or a character grouping variable to define your groups or design categories. The data must be sorted by these values. If not, use Data to sort your file on the grouping variable before using the ANOVA dialog box.

The Stats ANOVA procedure is for one-way designs only. For more complex designs, including unusual contrasts in one-way designs, see MGLH.

Tukey HSD post hoc tests are optional. They protect you from declaring pairs of means different when they could differ by chance.

- Select Stats/One-way ANOVA... from the Stats menu
- Select the variables to be analyzed from the left selection list
- Select a grouping variable from the right selection list
- Select Tukey HSD test
- Click OK



© 1989, SYSTAT, Inc.

Post hoc pairwise comparisons

Stats offers the Tukey HSD test for post hoc comparison of means. SYSTAT formerly offered Newman-Keuls and Duncan tests. Theoretical and Monte Carlo research has indicated recently that these tests do not offer their stated protection levels and therefore they have been dropped from SYSTAT. (Miller, 1985 and Einot and Gabriel, 1975).

The Tukey HSD test requires that the counts of cases per cell be equal. If you have unequal *n*'s, SYSTAT automatically adjusts by a harmonic mean *n*. This is called the Tukey-Kramer adjustment.

© 1989, SYSTAT, Inc.

Computation

Missing data

SYSTAT eliminates missing data before computing summary statistics for each variable separately. The number of missing cases deleted is be reported.

Algorithms

SYSTAT uses a one-pass provisional algorithm and double precision arithmetic for all computations (Spicer, 1972). Wilkinson and Dallal (1977) summarize the performance of this algorithm versus ones used in several statistical packages.

The Tukey test is computed from the exact probabilities for the upper quantiles of the Studentized Range (Lund & Lund, 1983). The algorithm used by SYSTAT is a recent one that takes considerable computer time.

© 1989, SYSTAT, Inc.

10.1 Basic statistics

Here is an example of basic statistics computed on selected variables in the MEDICAL data file. These are death rates from various illnesses for each of the 50 states.

- Select **Stats/Statistics...** from the **Stats** menu
- Select the variables to be analyzed: ACCIDENT, CARDIO, CANCER, PULMONAR, PNEU_FLU, DIABETES, LIVER
- Click OK

Total Observations: 50

	ACCIDENT	CARDIO	CANCER	PULMONAR	PNEU_FLU
N of cases	50	50	50	50	50
Minimum	28.600	114.600	76.100	8.300	12.400
Maximum	85.800	508.700	244.000	36.300	32.400
Mean	44.312	398.530	178.410	26.486	21.038
Standard Dev	10.983	84.244	33.051	5.602	4.140

	DIABETES	LIVER
N of cases	50	50
Minimum	3.400	6.200
Maximum	26.200	17.700
Mean	14.836	10.548
Standard Dev	3.668	2.745

10.2 Statistics by groups

Here is an example of the same analysis done by group. This output was produced by the following:

- Select **By groups...** from the **Data** menu
- Select the grouping variable: REGION (make sure the file is already sorted by REGION)
- Click OK

- Select **Stats/Statistics...** from the **Stats** menu
- Choose the statistics to be computed: [Mean], [SD]
- Click OK

© 1989, SYSTAT, Inc.

The following results are for:
REGION = 1.000

Total observations: 9

	ACCIDENT	CARDIO	CANCER	PULMONAR	PNEU_FLU
N of cases	9	9	9	9	9
Minimum	28.600	395.900	182.200	23.200	17.700
Maximum	41.300	508.700	231.500	33.600	29.700
Mean	34.278	458.511	209.944	27.333	22.333
Standard Dev	3.864	35.907	16.042	4.013	3.698

The following results are for:
REGION = 2.000

Total observations: 12

	ACCIDENT	CARDIO	CANCER	PULMONAR	PNEU_FLU
N of cases	12	12	12	12	12
Minimum	32.900	398.300	173.700	18.200	18.000
Maximum	53.000	495.200	211.100	32.300	32.400
Mean	39.442	449.133	188.300	26.783	24.075
Standard Dev	6.322	32.234	10.862	4.506	4.530

The following results are for:
REGION = 3.000

Total observations: 16

	ACCIDENT	CARDIO	CANCER	PULMONAR	PNEU_FLU
N of cases	16	16	16	16	16
Minimum	35.200	327.900	146.500	18.600	15.800
Maximum	62.500	502.700	244.000	35.200	24.500
Mean	47.638	414.069	183.681	25.313	19.400
Standard Dev	7.061	50.662	23.459	4.858	2.624

The following results are for:
REGION = 4.000

Total observations: 13

	ACCIDENT	CARDIO	CANCER	PULMONAR	PNEU_FLU
N of cases	13	13	13	13	13
Minimum	32.500	114.600	76.100	8.300	12.400
Maximum	85.800	387.800	179.400	36.300	25.100
Mean	51.662	291.169	140.962	27.069	19.354
Standard Dev	14.531	76.755	33.062	8.154	4.042

© 1989, SYSTAT, Inc.

10.3 Two-sample t-test

The following data show life expectancies for males and females for a sample of countries. We would like to see if average life expectancy differs for males and females.

SEX$	COUNTRY$	LIFE
Male	U.S.	71.0
Female	U.S.	78.3
Male	Canada	71.9
Female	Canada	78.9
Male	Mexico	62.1
Female	Mexico	66.0
Male	Brazil	57.6
Female	Brazil	61.1
Male	Chile	61.3
Female	Chile	70.6
Male	Venezuela	67.1
Female	Venezuela	70.7
Male	France	70.4
Female	France	78.5
Male	W. Germany	70.4
Female	W. Germany	77.0
Male	Hungary	65.5
Female	Hungary	73.4
Male	Italy	69.7
Female	Italy	75.9
Male	Norway	72.6
Female	Norway	79.5
Male	U.S.S.R.	64.0
Female	U.S.S.R.	74.0
Male	India	46.4
Female	India	44.7
Male	Iran	55.7
Female	Iran	55.0
Male	Japan	72.2
Female	Japan	79.7
Male	Kenya	51.2
Female	Kenya	54.7
Male	Australia	72.0
Female	Australia	78.7

For a two-sample *t*-test on these data:

- Select **Stats/t-test...** from the **Stats** menu
- Choose type of test to be performed: [Independent]
- Select the variable: LIFE
- Select the grouping variable: SEX$
- Click OK

© 1989, SYSTAT, Inc.


```
Independent samples t-test on     LIFE     grouped by      SEX$

    Group          N     Mean           SD
     Male         17     64.771         8.012
     Female       17     70.394        10.560

 Pooled variances t =       1.749 df =    32 prob = .090
```

Notice that, despite the difference in average life expectancy between 70.3 and 64.8 years, we cannot reject the null hypothesis that a difference this large could occur by chance.

If the number of cases differs between groups, you get two *t*-statistics: SEPARATE and POOLED. SEPARATE uses the separate variances within groups for its error. The degrees of freedom follow Brownlee (1965). POOLED uses the pooled within-groups variance, as in ANOVA. If the results differ substantially, you should probably use SEPARATE.

10.4 Dependent t-test

Now, let's do a dependent *t*-test on the same data. There is such great variability among countries that we should consider pairing males and females within country. This should improve the sensitivity of the test.

To do a dependent test, you need to have the values of both variables on the same case record. Here is our rearranged dataset:

COUNTRY$	MALE	FEMALE
U.S.	71.0	78.3
Canada	71.9	78.9
Mexico	62.1	66.0
Brazil	57.6	61.1
Chile	61.3	70.6
Venezuela	67.1	70.7
France	70.4	78.5
W. Germany	70.4	77.0
Hungary	65.5	73.4
Italy	69.7	75.9
Norway	72.6	79.5
U.S.S.R.	64.0	74.0
India	46.4	44.7
Iran	55.7	55.0
Japan	72.2	79.7
Kenya	51.2	54.7
Australia	72.0	78.7

© 1989, SYSTAT, Inc.

Here is how to get a paired test for these data (in NEWLIFE):

- Select **Stats/t-test...** from the **Stats** menu
- Choose type of test to be performed: [Paired]
- Select the variables: MALE, FEMALE
- Click OK

```
Paired Samples t-test on     MALE     vs    FEMALE     with     17 cases.

Mean difference =        -5.624
SD difference   =         3.221
DF      =        16
T       =         7.199
Prob    =         0.000
```

We lost half our degrees of freedom by pairing the values , but we got a lot of help by removing the between country variation. Now we find a highly significant difference in life expectancy between males and females.

10.5 Analysis of variance

The following data are iris flower measurements from Anderson (1935), reprinted in Fisher (1936). The measurements are sepal length, sepal width, petal length, and petal width in centimeters. Three species were measured: Setosa, Versicolor, and Virginica (corresponding to the indices 1, 2, and 3 in the data set).

The numbers in the following data set have been spread across three columns to save space. For input to SYSTAT, the three columns should be stacked vertically in one long column. The variables in the IRIS file are SPECIES, SEPALLEN, SEPALWID, PETALLEN, and PETALWID.

© 1989, SYSTAT, Inc.

1	5.1	3.5	1.4	0.2	2	7.0	3.2	4.7	1.4	3	6.3	3.3	6.0	2.5
1	4.9	3.0	1.4	0.2	2	6.4	3.2	4.5	1.5	3	5.8	2.7	5.1	1.9
1	4.7	3.2	1.3	0.2	2	6.9	3.1	4.9	1.5	3	7.1	3.0	5.9	2.1
1	4.6	3.1	1.5	0.2	2	5.5	2.3	4.0	1.3	3	6.3	2.9	5.6	1.8
1	5.0	3.6	1.4	0.2	2	6.5	2.8	4.6	1.5	3	6.5	3.0	5.8	2.2
1	5.4	3.9	1.7	0.4	2	5.7	2.8	4.5	1.3	3	7.6	3.0	6.6	2.1
1	4.6	3.4	1.4	0.3	2	6.3	3.3	4.7	1.6	3	4.9	2.5	4.5	1.7
1	5.0	3.4	1.5	0.2	2	4.9	2.4	3.3	1.0	3	7.3	2.9	6.3	1.8
1	4.4	2.9	1.4	0.2	2	6.6	2.9	4.6	1.3	3	6.7	2.5	5.8	1.8
1	4.9	3.1	1.5	0.1	2	5.2	2.7	3.9	1.4	3	7.2	3.6	6.1	2.5
1	5.4	3.7	1.5	0.2	2	5.0	2.0	3.5	1.0	3	6.5	3.2	5.1	2.0
1	4.8	3.4	1.6	0.2	2	5.9	3.0	4.2	1.5	3	6.4	2.7	5.3	1.9
1	4.8	3.0	1.4	0.1	2	6.0	2.2	4.0	1.0	3	6.8	3.0	5.5	2.1
1	4.3	3.0	1.1	0.1	2	6.1	2.9	4.7	1.4	3	5.7	2.5	5.0	2.0
1	5.8	4.0	1.2	0.2	2	5.6	2.9	3.6	1.3	3	5.8	2.8	5.1	2.4
1	5.7	4.4	1.5	0.4	2	6.7	3.1	4.4	1.4	3	6.4	3.2	5.3	2.3
1	5.4	3.9	1.3	0.4	2	5.6	3.0	4.5	1.5	3	6.5	3.0	5.5	1.8
1	5.1	3.5	1.4	0.3	2	5.8	2.7	4.1	1.0	3	7.7	3.8	6.7	2.2
1	5.7	3.8	1.7	0.3	2	6.2	2.2	4.5	1.5	3	7.7	2.6	6.9	2.3
1	5.1	3.8	1.5	0.3	2	5.6	2.5	3.9	1.1	3	6.0	2.2	5.0	1.5
1	5.4	3.4	1.7	0.2	2	5.9	3.2	4.8	1.8	3	6.9	3.2	5.7	2.3
1	5.1	3.7	1.5	0.4	2	6.1	2.8	4.0	1.3	3	5.6	2.8	4.9	2.0
1	4.6	3.6	1.0	0.2	2	6.3	2.5	4.9	1.5	3	7.7	2.8	6.7	2.0
1	5.1	3.3	1.7	0.5	2	6.1	2.8	4.7	1.2	3	6.3	2.7	4.9	1.8
1	4.8	3.4	1.9	0.2	2	6.4	2.9	4.3	1.3	3	6.7	3.3	5.7	2.1
1	5.0	3.0	1.6	0.2	2	6.6	3.0	4.4	1.4	3	7.2	3.2	6.0	1.8
1	5.0	3.4	1.6	0.4	2	6.8	2.8	4.8	1.4	3	6.2	2.8	4.8	1.8
1	5.2	3.5	1.5	0.2	2	6.7	3.0	5.0	1.7	3	6.1	3.0	4.9	1.8
1	5.2	3.4	1.4	0.2	2	6.0	2.9	4.5	1.5	3	6.4	2.8	5.6	2.1
1	4.7	3.2	1.6	0.2	2	5.7	2.6	3.5	1.0	3	7.2	3.0	5.8	1.6
1	4.8	3.1	1.6	0.2	2	5.5	2.4	3.8	1.1	3	7.4	2.8	6.1	1.9
1	5.4	3.4	1.5	0.4	2	5.5	2.4	3.7	1.0	3	7.9	3.8	6.4	2.0
1	5.2	4.1	1.5	0.1	2	5.8	2.7	3.9	1.2	3	6.4	2.8	5.6	2.2
1	5.5	4.2	1.4	0.2	2	6.0	2.7	5.1	1.6	3	6.3	2.8	5.1	1.5
1	4.9	3.1	1.5	0.2	2	5.4	3.0	4.5	1.5	3	6.1	2.6	5.6	1.4
1	5.0	3.2	1.2	0.2	2	6.0	3.4	4.5	1.6	3	7.7	3.0	6.1	2.3
1	5.5	3.5	1.3	0.2	2	6.7	3.1	4.7	1.5	3	6.3	3.4	5.6	2.4
1	4.9	3.6	1.4	0.1	2	6.3	2.3	4.4	1.3	3	6.4	3.1	5.5	1.8
1	4.4	3.0	1.3	0.2	2	5.6	3.0	4.1	1.3	3	6.0	3.0	4.8	1.8
1	5.1	3.4	1.5	0.2	2	5.5	2.5	4.0	1.3	3	6.9	3.1	5.4	2.1
1	5.0	3.5	1.3	0.3	2	5.5	2.6	4.4	1.2	3	6.7	3.1	5.6	2.4
1	4.5	2.3	1.3	0.3	2	6.1	3.0	4.6	1.4	3	6.9	3.1	5.1	2.3
1	4.4	3.2	1.3	0.2	2	5.8	2.6	4.0	1.2	3	5.8	2.7	5.1	1.9
1	5.0	3.5	1.6	0.6	2	5.0	2.3	3.3	1.0	3	6.8	3.2	5.9	2.3
1	5.1	3.8	1.9	0.4	2	5.6	2.7	4.2	1.3	3	6.7	3.3	5.7	2.5
1	4.8	3.0	1.4	0.3	2	5.7	3.0	4.2	1.2	3	6.7	3.0	5.2	2.3
1	5.1	3.8	1.6	0.2	2	5.7	2.9	4.2	1.3	3	6.3	2.5	5.0	1.9
1	4.6	3.2	1.4	0.2	2	6.2	2.9	4.3	1.3	3	6.5	3.0	5.2	2.0
1	5.3	3.7	1.5	0.2	2	5.1	2.5	3.0	1.1	3	6.2	3.4	5.4	2.3
1	5.0	3.3	1.4	0.2	2	5.7	2.8	4.1	1.3	3	5.9	3.0	5.1	1.8

© 1989, SYSTAT, Inc.

Next we produce a one-way analysis of variance on the four measurements across the three species. The options specify a post-hoc test on pairs of means using Tukey's HSD procedure. Note that the cases are already in order by SPECIES and do not have to be sorted.

- Select **Stats/One-way ANOVA...** from the **Stats** menu
- Select the dependent variables:
  SEPALLEN,SEPALWID,PETALLEN,PETALWID
- Select the grouping variable: SPECIES
- Click [Tukey HSD test]
- Click OK

```
The Following Results are for:
        SPECIES     =        1.000

Total Observations:   50

                  SPECIES    SEPALLEN    SEPALWID    PETALLEN    PETALWID

  N of Cases           50          50          50          50          50
  Mean              1.000       5.006       3.428       1.462       0.246
  Standard Dev      0.000       0.352       0.379       0.174       0.105

The Following Results are for:
        SPECIES     =        2.000

Total Observations:   50

                  SPECIES    SEPALLEN    SEPALWID    PETALLEN    PETALWID

  N of Cases           50          50          50          50          50
  Mean              2.000       5.936       2.770       4.260       1.326
  Standard Dev      0.000       0.516       0.314       0.470       0.198

The Following Results are for:
        SPECIES     =        3.000

Total Observations:   50

                  SPECIES    SEPALLEN    SEPALWID    PETALLEN    PETALWID

  N of Cases           50          50          50          50          50
  Mean              3.000       6.588       2.974       5.552       2.026
  Standard Dev      0.000       0.636       0.322       0.552       0.275

Summary Statistics for SEPALLEN

Bartlett Test for Homogeneity of Group Variances

Chi-Square=      16.006 DF=  2 Probability =  .000

                        Analysis of Variance

Source          Sum of Squares  DF  Mean SQUARE        F    Probability

Between Groups          63.212   2       31.606    119.265       .000
Within Groups           38.956 147        0.265
```


© 1989, SYSTAT, Inc.


```
Matrix of Pairwise Absolute Mean Differences

                1           2           3

  1       0.000
  2       0.930       0.000
  3       1.582       0.652       0.000

Tukey HSD Multiple Comparisons
Matrix of Pairwise Comparison Probabilities

                1           2           3

  1       1.000
  2       0.000       1.000
  3       0.000       0.002       1.000

Summary Statistics for SEPALWID

Bartlett Test for Homogeneity of Group Variances

Chi-Square=       2.091 DF=  2 Probability =  .352

                        Analysis of Variance

Source          Sum of Squares  DF  Mean SQUARE        F    Probability

Between Groups          11.345   2         5.672     49.160       .000
Within Groups           16.962 147         0.115

Matrix of Pairwise Absolute Mean Differences

                1           2           3

  1       0.000
  2       0.658       0.000
  3       0.454       0.204       0.000

Tukey HSD Multiple Comparisons
Matrix of Pairwise Comparison Probabilities

                1           2           3

  1       1.000
  2       0.000       1.000
  3       0.002       0.008       1.000

Summary Statistics for PETTALLEN

Bartlett Test for Homogeneity of Group Variances

Chi-Square=      55.423 DF=  2 Probability =  .000

                        Analysis of Variance

Source          Sum of Squares  DF  Mean SQUARE        F    Probability

Between Groups         437.103   2       218.551   1180.161       .000
Within Groups           27.223 147         0.185
```

© 1989, SYSTAT, Inc.

```
Matrix of Pairwise Absolute Mean Differences

                    1           2           3

          1     0.000
          2     2.798       0.000
          3     4.090       1.292       0.000

Tukey HSD Multiple Comparisons
Matrix of Pairwise Comparison Probabilities

                    1           2           3

          1     1.000
          2     0.000       1.000
          3     0.000       0.000       1.000

Summary Statistics for PETALWID

Bartlett Test for Homogeneity of Group Variances

Chi-Square=      39.213 DF=  2 Probability =  .000

                          Analysis of Variance

Source          Sum of Squares  DF  Mean SQUARE        F    Probability

Between Groups          80.413   2       40.207     960.007      .000
Within Groups            6.157 147        0.042

Matrix of Pairwise Absolute Mean Differences

                    1           2           3

          1     0.000
          2     1.080       0.000
          3     1.780       0.700       0.000

Tukey HSD Multiple Comparisons
Matrix of Pairwise Comparison Probabilities

                    1           2           3

          1     1.000
          2     0.000       1.000
          3     0.000       0.000       1.000
```

SYSTAT produces a test for homogeneity of variances plus an ANOVA for each variable. Because the cell counts are greater than 10, SYSTAT uses the Bartlett chi-square. If the cell counts are less than 10, Box's small sample *F* approximation is used for the probability value rather than Bartlett's usual chi-square (Dixon and Massey, 1969, p. 308).

© 1989, SYSTAT, Inc.

For the Tukey HSD test, the absolute value of the difference between means is printed for each pair of variables in a triangular matrix. The rows/columns of this matrix are numbered with indices of the groups in the same order that they appear in your file. The same format is used for the associated probabilities. Notice that every pairwise comparison is highly significant on every variable. If you want to print more decimal places in the probabilities, use **Data/Formats**.

© 1989, SYSTAT, Inc.

11 Tables

Introduction 472
Making tables 472
One-way tables 472
Two-way tables 473
Standardizing tables with percents 473
Significance tests 475
One-way tables 475
Two-way tables 477
McNemar's test 480
Multi-way tables 480
Measures of association for two-way tables 480
Phi, Cramer's V, and contingency 481
Cohen's kappa 481
Ordinal association 481
Measures for 2×2 tables 481
Log-linear models 481
Usage 483
Tabulate... 483
One-way tables 484
Two-way tables 484
Multi-way tables 484
Options 485
Tests and measures of association for two-way tables 486
Large tables 486
Log-linear models... 487
Options 489
Saving tables 490
Computation 491
Missing data 491
Algorithms 491
Examples 492
11.1 One-way frequency table 492
11.2 List table 492
11.3 Confidence intervals 493
11.4 Two-way table with tests and association measures 494
11.5 Percents 496
11.6 Mantel-Haenszel test 497
11.7 Log-linear model 499
11.8 Log-linear model of multi-way tables 500

© 1989, SYSTAT, Inc.

Tables 11

Overview

When variables are categorical, frequency tables (cross-tabulations) provide useful summaries. Sometimes, for a report, you may need only the number of cases falling in specified categories or cross-classifications. Other times you may require a test of independence or a measure of association between two categorical variables. Or you may want to model relationships among two or more categorical variables by fitting a log-linear model to the cell frequencies.

Tables can make, analyze, and save frequency tables that are formed by one, two, or more categorical variables. The values of the categorical variables (or table factors) may be character or numeric. Categories can be unordered (male and female, the four regions of the U.S.), ordered (number of buses with engine failures, number of children), or intervals defined on continuous variables (age split into child, teen, and adult). **Tables** forms tables using data read from a cases-by-variables rectangular file or data recorded as frequencies (from a table in a report) with cell indices.

Other procedures in SYSTAT compute statistics on categorical data. **Corr**, for example, can calculate gamma and tau-b coefficients. **Stats** can create files of summary data for tabular presentation. **Bar**, **Pie**, and **Cplot** can create graphics of categorical data.

© 1989, SYSTAT, Inc.

Introduction

Tables organize data into arrays. The simplest form of table is a matrix of numbers or characters. The most complex can be as extensive as the U.S. Budget or a census. What distinguishes tables from other modes of organizing data is that tables are stratified on categorical variables. Unless you have a variable that contains exclusive categories, you cannot produce a table.

Making tables

There are many formats for displaying tabular data. Let's examine basic layouts for counts and percents.

One-way tables

Here is an example of a table showing population in thousands categorized by sex from the 1980 U.S. Census.

```
Table of values for      SEX$

Frequencies

        female      male        TOTAL
-------------------------------------
        122579      116160      238739
-------------------------------------
```

The categorical variable producing this table is SEX$, or gender. The entries of the table are counts. Table entries can also be budget figures, statistics such as standard deviations, and even character data such as names and addresses. SYSTAT tabulates counts.

Sometimes, you may define categories as intervals of a continuous variable. Here is an example showing the U.S. population broken down by age.

```
Table of values for      AGE$

Frequencies

    0-9       10-19     20-29     30-39     40-49     50-59
-------------------------------------------------------------
    34857     35654     42745     37975     25703     22278
-------------------------------------------------------------

    60-69     70-79     80-       TOTAL
-------------------------------------------
    20427     13060     6040      238739
-------------------------------------------
```


© 1989, SYSTAT, Inc.

When you categorize a continuous variable, it is tempting to make even cutpoints, as in this example. Sometimes, however, you should consider uneven cutpoints in order to follow natural or logical separations in your data.

Two-way tables

A *cross-tabulation* is a table that displays one cell for every combination of values on two or more categorical variables. Here is a two-way table that crosses the age and sex distributions of the tables above.

```
Table of      AGE$     (rows) BY      SEX$     (columns)

Frequencies

          female      male        TOTAL
-------------------------------------
0-9       17019       17838       34857
10-19     17447       18207       35654
20-29     21344       21401       42745
30-39     19138       18837       37975
40-49     13135       12568       25703
50-59     11617       10661       22278
60-69     11053        9374       20427
70-79      7712        5348       13060
80-        4114        1926        6040
-------------------------------------
TOTAL     122579      116160      238739
-------------------------------------
```

This cross-tabulation shows relationships between age and sex, which were invisible in the separate tables. Notice, for example, that older males decline in population relative to females.

Standardizing tables with percents

As with other statistical procedures such as correlation, it sometimes helps to have numbers standardized on a recognizable scale. Correlations vary between –1 and 1, for example. A convenient scale for table counts is percentage, which varies between 0 and 100.

With tables, you must choose a facet on which to standardize: rows, columns, or the total count in the table. For example, if we are interested in looking at the difference between male and female population within age groups, we might want to standardize by rows, so that we can see the percent of males and females within each age group. Here is that table:

© 1989, SYSTAT, Inc.

Table of AGE$ (rows) BY SEX$ (columns)

Row percents

	female	male	TOTAL
0-9	48.83	51.17	100.00
10-19	48.93	51.07	100.00
20-29	49.93	50.07	100.00
30-39	50.40	49.60	100.00
40-49	51.10	48.90	100.00
50-59	52.15	47.85	100.00
60-69	54.11	45.89	100.00
70-79	59.05	40.95	100.00
80-	68.11	31.89	100.00
TOTAL	51.34	48.66	100.00

The news looks grim for males (or females, depending on your point of view).

On the other hand, if we are interested in the overall distribution of males vs. females, we might want to standardize within columns to compare the age distributions of males and females:

Table of AGE$ (rows) BY SEX$ (columns)

Column percents

	female	male	TOTAL
0-9	13.88	15.36	14.60
10-19	14.23	15.67	14.93
20-29	17.41	18.42	17.90
30-39	15.61	16.22	15.91
40-49	10.72	10.82	10.77
50-59	9.48	9.18	9.33
60-69	9.02	8.07	8.56
70-79	6.29	4.60	5.47
80-	3.36	1.66	2.53
TOTAL	100.00	100.00	100.00

Notice that, overall, the distribution of age is skewed, with fewer old people.

© 1989, SYSTAT, Inc.

Significance tests

After producing a table, you may wish to consider a population model that accounts for the structure you see in the observed table. As we noted in the **Introduction to statistics** chapter, you should have a population in mind when you make such inferences. Since our data here *are* the population of the U.S. in 1980, we must view it as a sample from one point in time. That doesn't necessarily make it a simple random sample, but we are going to overlook that technicality. You should keep in mind, however, that many published statistical analyses of tables do not explicitly deal with the sampling problem.

One-way tables

A model for our data might be that the proportion of the males and females is equal in the population. The null hypothesis corresponding to the model is:

$$H_0: p_{males} = p_{females}$$

The sampling model for testing this hypothesis requires that a population contains equal numbers of males and females and that each member of the population has an equal chance of being chosen. After choosing each person, we identify the person as male or female. There is no other category possible and one person cannot fit under both categories (*exhaustive* and *mutually exclusive*).

There is an exact way to reject our null hypothesis (called a *permutation test*). We can tally every possible sample of size 238739 (including one with no females and one with no males). Then we can sort our samples into two piles: samples in which there are between 48.66 and 51.34 percent females and samples in which there are not. If the latter pile is extremely small relative to the former, we can reject the null hypothesis.

Needless to say, this would be a tedious undertaking—particularly on a microcomputer. Fortunately, there is an approximation using a continuous probability distribution that works quite well. First we need to calculate the expected count of males and females, respectively, in a sample of size 238739 if p is .5. This is 119369.5, or half the sample N. Next, we subtract the observed counts from these expected counts, square them, and divide by the expected:

$$X^2 = \frac{(122579-119369.5)^2}{119369.5} + \frac{(116160-119369.5)^2}{119369.5} = 172.59$$

© 1989, SYSTAT, Inc.

If our assumptions about the population and the structure of the table are correct, then this statistic will be distributed as a mathematical chi-square variable. We can look up the area under the tail of the chi-square statistic beyond the sample value we calculate and if this area is small (say, less than .05), we can reject the null hypothesis.

To look up the value, we need a *degrees of freedom* (*df*) value. This is the number of independent values being added together to produce the chi-square. In our case, it is 1, since the observed proportion of men is simply 1 minus the observed proportion of women. If there were three categories (men, women, other?), then the degrees of freedom would be 2. Anyway, if you look up the value 172.59 with one degree of freedom in your chi-square table, you will find the probability of exceeding this value is exceedingly small. Thus, we reject our null hypothesis that the proportion of males equals the proportion of females in the population.

This chi-square approximation is good only for large samples. A popular rule of thumb is that the expected counts should be greater than 5, although it should be even greater if you wish to be comfortable with your test. With our sample, the difference between the approximation and the exact result is negligible. For both, the probability is tiny.

Have you detected the flaw in our reasoning? We seem to have forgotten that the counts of males and females are in thousands. Therefore, our chi-square statistic is really a thousand times larger and the result is even more significant. If your sample is large enough, even tiny differences will be significant.

Our hypothesis test has an associated confidence interval. You can use SYSTAT to compute this interval on the population data. This time, we multiplied the counts by 1000 before computing the statistics. Here is the result:

```
95 percent approximate confidence intervals scaled as cell percents

   female      male
-----------------------
    51.35     48.66
    51.34     48.65
-----------------------
```

The confidence intervals vary only in the last digit (lower is on the bottom, upper on the top). The standard errors of these percentages are less than .01.

© 1989, SYSTAT, Inc.

In conclusion, remember that there are methodological problems in making inferences like these when the data are not a simple random sample from a well defined population. We have gone through this example to illustrate some of these pitfalls and because of them, we would not publish these results.

Two-way tables

The chi-square procedure for testing the null hypothesis of specific population proportions in a one-way table can be extended to hypotheses about multi-way tables. For two-way tables, the most popular null hypothesis is additive. That is, there is no interaction in the table.

The concept of interaction in a two-way frequency table is similar to the one in analysis of variance. It is easiest to see in an example. Schachter (1959) randomly assigned 30 subjects to one of two groups: High Anxiety (17 subjects), who were told that they would be experiencing painful shocks, and Low Anxiety (13 subjects), who were told that they would experience painless shocks. After the assignment, each subject was given the choice of waiting alone or with the other subjects. The following tables illustrate two possible outcomes of this study.

No Interaction

		WAIT: Alone	WAIT: Together
ANXIETY	High	8	9
	Low	6	7

Interaction

		WAIT: Alone	WAIT: Together
ANXIETY	High	5	12
	Low	9	4

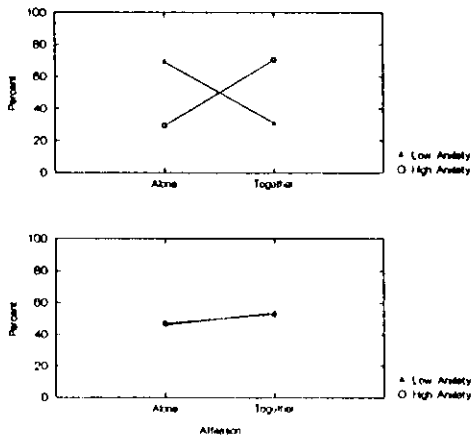
Notice in the table on the left that the number choosing to wait together relative to those choosing to wait alone is similar for both High and Low Anxiety groups. In the right table, on the other hand, more of the High Anxiety group chose to wait together.

© 1989, SYSTAT, Inc.

We are interpreting these numbers relatively, so we should compute row percents to understand the differences better. Here are the same tables standardized by rows:

		No Interaction WAIT		*Interaction* WAIT	
		Alone	Together	Alone	Together
ANXIETY	High	47.1	52.8	29.4	70.6
	Low	46.1	53.8	69.2	30.8

Now we can see that the percentages are similar in the two rows of the left table (No Interaction) and quite different in the right (Interaction). A simple graph reveals these differences even more strongly. In the following figure, the No Interaction row percents are plotted below the Interaction percents.



Notice that the lines cross in the upper plot, showing that the rows differ. There is almost complete overlap in the lower plot.

© 1989, SYSTAT, Inc.

Now, in the one-way table example above, we tested the hypothesis that the cell proportions were equal in the population. We can test an analogous hypothesis in this context—that each of the four cells contains 25 percent of the population. The problem with this thinking is that we know already that Schachter assigned randomly more people to the High Anxiety group. In other words, the row marginal percentages (or totals) we should take as fixed when we determine what proportions to expect in the cells from a random model.

Our No Interaction model is based on these fixed marginals. In fact, we can fix either the row or column margins to compute a No Interaction model, because the total number of subjects is fixed at 30. You can verify that the row and column sums in the above tables are the same.

Now we are ready to compute our chi-square test of interaction (often called *test of independence*) in the two-way table by using the No Interaction counts as expected counts in our chi-square formula above. This time, our degrees of freedom are still 1, because the marginal counts are fixed. If you know the marginal counts, then one cell count determines the remaining three. In general, the degrees of freedom for this test are (rows–1) times (columns–1).

Here is the result of our chi-square test. The chi-square is 4.693, with a p of .03. On this basis, we reject our No Interaction hypothesis.

```
Table of ANXIETY$  (rows) BY  WAIT$  (columns)

Frequencies

          Alone   Together    TOTAL
High          5         12       17
Low           9          4       13
TOTAL        14         16       30

 Test statistic                 Value        DF          Prob
  Pearson Chi-square            4.693         1          .030
  Likelihood ratio chi-square   4.810         1          .028
  McNemar Symmetry Chi-square    .429         1          .513
  Yates Corrected Chi-square    3.229         1          .072
  Fisher Exact Test (Two-tail)                             .063
```

© 1989, SYSTAT, Inc.

Actually, we cheated. The program computed the expected counts from the observed data. These are not exactly the ones we showed you in the No Interaction table. They differ by rounding error in the first decimal place. You can compute them exactly if you wish. The popular method is to multiply the total row count times the total column count corresponding to a cell and dividing by the total sample size. For the upper left cell, this would be 17*14/30 = 7.93

There is one other interesting problem with these data. Remember when we mentioned that the chi-square is only an approximation and that it does not work well for small samples. Although these data meet the minimum expected count of 5 criterion, they are nevertheless problematic. Look at the Fisher's exact test result in the output above. Like our permutation test above, which was so cumbersome for large datasets, Fisher's test counts all possible outcomes exactly, including the ones that produce interaction greater than what we observed. The Fisher exact test *p* value is not significant (.063). On this basis, we could not reject the null hypothesis of no interaction, or independence.

Yates' chi-square test in the output is an attempt to adjust the Pearson chi-square statistic for small samples. While it has come into disfavor for being unnecessarily conservative in many instances, nevertheless, the Yates *p* value is consistent with Fisher's in this case (.072).

McNemar's test

For two-way square tables, in which the data represent matched pairs or paired comparisons, you should consider McNemar's test. The structure it tests arises when the same subjects are measured twice, as in a paired comparisons *t* test, or when subjects are paired or matched. This test ignores the counts along the diagonal of the table and tests if the counts in cells above the diagonal differ from those below the diagonal.

Multi-way tables

It is possible to extend the test of independence to multi-way tables by computing expected counts with a series of marginal multiplications in a similar manner. For more detailed hypotheses, however, you will probably want to use log-linear modeling (see below).

Measures of association for two-way tables

For two-way tables there are many indices that measure association between rows and columns. Some of these standardize, like the Pearson correlation, the degree of interaction that we tested for significance with the Pearson chi-square. Others measure different types of association, such as symmetry in square tables. We will summarize some of these statistics below. See Example 11.4 for sample output.

© 1989, SYSTAT, Inc.

Phi, Cramer's V, and contingency

These measures all quantify the degree of interaction tested by Pearson's chi-square. In fact, all are based on the chi-square formula. For 2×2 tables, phi is scaled the same as a Pearson correlation computed on binary variables.

Cohen's kappa

The table structure for Cohen's kappa looks like that for McNemar's test in that the row and column categories are the same. But here the focus shifts to the diagonal: are the counts along the diagonal significantly greater than those expected by chance alone? Because each subject is classified or rated twice, kappa is a measure of the interrater agreement. Its size is judged by using an asymptotic standard error to construct a confidence interval. Values of kappa greater than .75 have been said to indicate strong agreement; between .40 and .79 fair to good, and below .40, poor agreement.

Ordinal association

Kendall's tau-b, Stuart's tau-c, Goodman-Kruskal gamma, and Spearman's rho are appropriate when both categorical variables in a two-way table have ordered categories. The first three measures differ only in how ties are treated and the fourth is like the usual Pearson correlation except that the rank order of each value is used in the computations instead of the value itself.

Measures for 2×2 tables

Yule's Q and Y measure dominance in a 2×2 table. That is, both statistics are 1 if either off-diagonal cell is zero. Otherwise, they are less than 1 in absolute value. See **Corr** for additional association measures for binary data (S2–S6, provided by **Dichotomous...**).

Log-linear models

Log-linear models represent relations among categorical variables in a contingency table by linear effects on logarithms of observed cell frequencies. In a log-linear model, a main effect for a factor (categorical variable) represents unequal marginal frequencies for the levels of the factor. An interaction term represents the association between categorical variables—the joint effect of these variables on cell frequencies. For more information on log-linear models, consult Cox (1970), Bishop, Fienberg, and Holland (1975), Kritzer (1979), or Fienberg (1980).

© 1989, SYSTAT, Inc.

A *fully saturated* log-linear model looks like a complete factorial analysis of variance model, except the parameters are in logarithmic units. The log-linear model expresses the log of the count in each cell as a function of main effect and interaction parameters. The two methods differ in focus, however. In analysis of variance, we usually seek significant main effects and assume interactions are nuisances. In log linear modeling, we usually seek significant interactions. Of course, there are exceptions to this rule.

The fully-specified model in the "Usage" section using the DANISH dataset directly corresponds to the models one sees in publications on log-linear analyses. Grant Blank has summarized this notation for a fully saturated model as follows:

SYSTAT	FATHER + SON + FATHER*SON *or* FATHER*SON
Agresti (1984)	$\log m_{ij} = \mu + \lambda_j^F + \lambda_j^S + \lambda_i^{FS}_j$
Fienberg (1980)	$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{1\,2(ij)}$
Goodman (1978)	$\xi_{ij} = \theta + \lambda_i^A + \lambda_j^B + \lambda_i^A{}_j^B$
Haberman (1978)	$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_i^A{}_j^B$
Knoke & Burke (1980)	$G_{ij} = \theta + \lambda_j^F + \lambda_j^S + \lambda_i^{FS}_j$
or, in multiplicative form, Goodman (1971)	$F_{ij} = \eta r_i^A r_j^B r_{ij}^{AB}$ where $\xi_{ij} = \log(F_{ij})$, $\theta = \log \eta$, $\lambda_i^A = \log(r_i^A)$, etc.

Fitting this fully saturated model is pointless since the fitted values equal the observed values and the chi-square badness-of-fit statistics are zero. The usual model for a two-way table like this leaves out the interaction (the association of the two variables) to see how well an additive model works. For a two-way table, testing this model with the Pearson chi-square is equivalent to the customary test of independence:

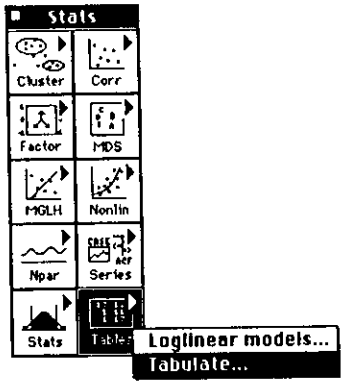
```
CONSTANT + FATHER + SON
```

This model is computed as Example 11.7.

© 1989, SYSTAT, Inc.

Usage

Tables has two items. **Tabulate...** provides one-way, two-way, and multi-way tables as well as tests of significance and measures of association. After using **Tabulate...**, you may fit a log-linear model to the table with **Log-linear models...**, which also computes chi-square and likelihood ratio statistics.



Press and hold the mouse button on **Tables**. A submenu pops up offering **Log-linear models...** and **Tabulate**. Select the item you want and release the mouse button.

Tabulate....

You can produce a variety of table formats with the **Tabulate...** dialog box. Options are available for percents and other information in table cells. If you select [Show extended results] in **Data/Formats...**, you can also get a variety of tests and measures of association for two-way tables.

© 1989, SYSTAT, Inc.

One-way tables

A one-way table is a horizontal display of frequencies for each value of a single categorical variable. To produce this display, select a variable from the left selection list. If you select more than one variable from the left selection list, you will get a separate display for each variable.

- Select **Tables/Tabulate...** from the **Stats** menu
  - Select one or more variables from the left selection list; all numeric variables are used by default
- Click OK

Two-way tables

A two-way table displays frequencies for each combination of values of two categorical variables in a row-by-column format. To produce a two-way table, select the row variable from the left selection list. Then select the column variable from the right. If you select more than one variable from the left selection list, you will get a separate two-way table for each row variable selected.

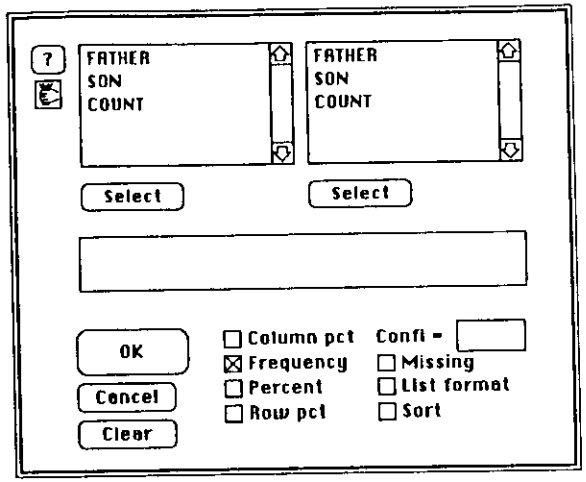
- Select **Tables/Tabulate...** from the **Stats** menu
  - Select one or more row variables from the left selection list
  - Select a column variable from the right selection list
- Click OK

Multi-way tables

Cells in a multi-way table are defined by all combinations of the values of three or more variables. SYSTAT prints these more complicated tables as a set of two-way slices through the higher dimensional classification. To produce multi-way tables, select a stratifying variable from the left selection list. Then select subsequent variables from the right selection list. The last two variables you select from the right selection list will be the row and column variables, respectively. If you select more than one variable from the left selection list, you will get a separate multi-way table for each selected variable.

- Select **Tables/Tabulate...** from the **Stats** menu
  - Select one or more stratifying variables from the left selection list
  - Select additional stratifying and row and column variables from the right selection list.
- Click OK

© 1989, SYSTAT, Inc.



Options

Percents

You can produce row, column, and total percentage tables. Normally, frequency tables are produced. If an option is selected, however, that option is used. The options are [Frequency] (cell count), [Percent] (percent of total table count), [Row] (percent of row total), and [Column] (percent of column total).

List output

The [List] option produces a table in list arrangement rather than the usual row-by-column arrangement. Statistics for every cell are printed on a separate line underneath each other. These include the frequency, the cumulative frequency, the percent of total and cumulative percent of total. This option is handy for producing a listing of the frequency and cumulative statistics for each distinct value of a continuous variable for data cleaning purposes.

Confidence intervals

For binomially or multinomially distributed data, you can get confidence intervals on cell proportions. Specify the confidence interval you want (e.g. .95) for [Confi]. This option works only for one-way tables.

© 1989, SYSTAT, Inc.

Missing values
Click [Missing] to exclude missing values from calculations. Otherwise, missing values will appear as an additional category for any variable containing them. Do not click this box if you wish to see the patterns of missing values in your file.

Sorting
When tabulating a variable, SYSTAT sorts the category values alphabetically or numerically. Because of this, the [Sort] option is normally checked. If you do not want the values sorted, turn off this option. When you do so, SYSTAT arranges the values in the order they appear in the data file. If you require a specific ordering of values, such as Low, Medium, High, you can make the first three cases in the file have this ordering.

Decimal places
For numeric category values, the Tables item prints three digits after the decimal point. [Decimals to be shown] in **Data/Formats...** can change this format. You can choose values from 0 (no decimal point) to 9 (9 digits after the decimal point). **Data/Formats...** does not affect character values or numbers printed inside the cells or the marginal totals and percents.

Tests and measures of association for two-way tables

If you choose [Show extended results] from **Data/Formats...**, SYSTAT prints significance tests, their associated coefficients, and Pearson's chi-square below two-way tables. For 2×2 tables, Yates' corrected chi-square and Fisher's exact test are printed. Coefficients include Phi, Cramer's V, Contingency, Asymmetric Lambda, Cohen's kappa, and Yule's Q and Y, where appropriate.

Large tables

SYSTAT stores tables data in computer memory. If you receive an "out of memory" message, you can still produce separate two-way tables by using **Data/By groups...** Suppose that you run out of memory tabulating A*B*C*D*E*F*G. You can sort the file on the first four variables, select those variables with **By groups...**, and tabulate the remaining three:

- Select **Sort...** from the **Data** menu
- Select sorting variables: A, B, C, D
- Click OK

© 1989, SYSTAT, Inc.

- Select **Open...** from the **File** menu
- Select the sorted file
- Click Open

- Select **By groups...** from the **Data** menu
- Select the grouping variables: A, B, C, D
- Click OK

- Select **Stats/Tabulate...** from the **Stats** menu
- Select variables: E, F, G
- Click OK

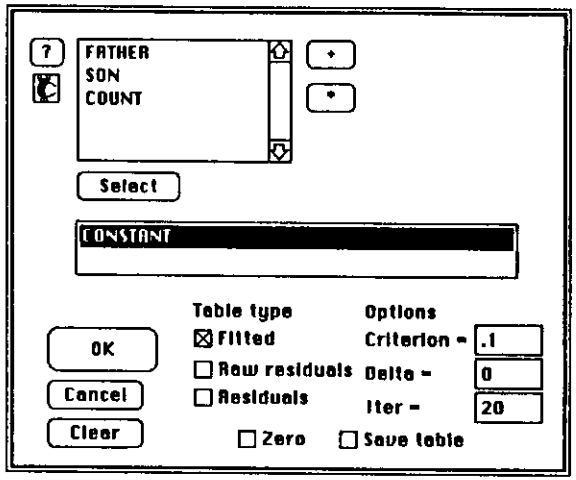
You should rarely run out of memory with **Tables**, but this approach can handle enormous problems if you tabulate everything with everything.

Log-linear models...

Tables/Log-linear models... fits a log-linear model to a table you have created with a prior **Tabulate...** procedure. (You cannot use **Log-linear models...** until you have used **Tabulate...**) Pearson chi-square and likelihood ratio statistics and an index of dissimilarity are shown.

- Select **Tables/Log-linear models...** from the **Stats** menu
  - Select individual variables from the selection list
  - Select a + or * to separate individual variables as appropriate
- Click OK

© 1989, SYSTAT, Inc.



Selecting variables

The model statement closely resembles that in MGLH, but is somewhat simpler; since the dependent variable is always the log of expected cell frequencies, it is omitted from the command. Also, all models are hierarchical—that is, all lower order effects are included in a model if higher order (interaction) effects are present.

For example, a fully saturated model that perfectly fits the observed frequencies for the DANISH data (listed in Example 11.4) is:

```
CONSTANT + FATHER*SON
```

Since the FATHER and SON main effects are implied by the interaction, all the following models are equivalent:

```
CONSTANT + FATHER + SON + FATHER*SON
CONSTANT + FATHER + FATHER*SON
CONSTANT + FATHER*SON + SON
CONSTANT + FATHER*SON
```


© 1989, SYSTAT, Inc.

Options

Diagnostic information

You can choose a table of [Fitted] values, [Raw residuals] or standardized [Residuals], computed from your log-linear model. [Fitted values] are the default.

Iterations

The fitting algorithm is iterative. Normally only a few iterations are required. The maximum number allowed is 20. If you wish to increase this number, use the [Iterations] option.

Criterion

Iterations terminate when either the maximum number is reached or the maximum difference between a fitted and observed value in a marginal table is less than .1. You can change this criterion with the [Criterion] option.

Delta

Sometimes tables have sparse cells. If a fitted value is exactly zero, it cannot be used in the chi-square statistics. In this case, the degrees of freedom are decremented for each zero fitted value. (Sometimes this makes the degrees of freedom too small; see Bishop, Fienberg, and Holland (1975).) SYSTAT warns if more than 20 percent of the cells have fitted values less than 5. To correct for this problem, either collapse categories or add a small constant to the cells with the [Delta] option. Many statisticians feel that this is a dangerous remedy for sparse tables. Do not use it unless you can replicate your results.

Structural zeros

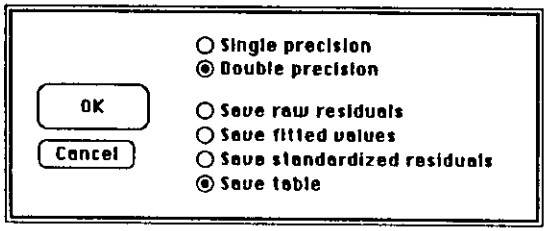
The term "structural zero" applies to cells that are empty naturally (e.g. pregnant males) or empty by design (e.g. excluding the diagonal of a two-way table). The term does not apply to cells that are empty because too few observations were collected. By clicking [Zero], you are saying that the probability is zero that an observation could occur in cells with zero observed frequencies.

The [Zero] test is done before adding [Delta] if you click both options. If you have cells with zero values that you do not want to treat as structural zeros, add one observation to all the cells except those to be held as structural zeros and then use both [Delta=-1] and [Zero].

© 1989, SYSTAT, Inc.

Saving tables

You can save a table into a SYSTAT file with the [Save table] option in the **Log-linear models...** dialog box. After you specify a filename and click [Save], a dialog box asks what type of table to save (fitted values, differences, standardized residuals, or cell counts) and whether to save in single or double precision.



Only one type of table may be saved into one file at a time. If you want to save both fitted values and raw residuals, for example, use **Log-linear models...** twice, clicking [Save tables] both times. Be sure to choose different filenames so the second table doesn't wipe out the first.

© 1989, SYSTAT, Inc.

Computation

Missing data

Missing data values are treated as an additional category for each variable. To prevent this, click [Missing] and the **Tables/Tabulate...** dialog box.

Algorithms

All computations are in double precision. The data are stored in memory, so there is an upper limit on the number of cells. If a two- or higher-way table is created, the maximum number of levels for any way is 50. **Tabulate...** works on one pass through the data, regardless of the number of tables requested.

Log-linear models... implements the algorithms of Haberman (1972). Storage is reallocated for the log-linear modeling stage after a table has been computed, so it may not be possible to estimate some models for large tables produced by **Tabulate...**

© 1989, SYSTAT, Inc

Examples

11.1 One-way frequency table

Tables's simplest use is to list frequencies of a variable. Here we use the US dataset. We are going to tabulate how many states there are in each census region.

- Select **Tables/Tabulate...** from the **Stats** menu
- Select a variable: REGION
- Click OK

```
Table of Values for   REGION

Frequencies
```

1	2	3	4	Total
9	12	16	13	50

Tables makes a separate cell for each distinct value of a variable. A variable with many distinct values may cause an out of memory error. SYSTAT tabulates missing data values as a separate category unless you click the [Ignore missing values] option. There are no missing data in REGION, so this makes no difference for our example.

11.2 List table

The [List format] option gives more information in a different format:

- Select **Tables/Tabulate...** from the **Stats** menu
- Choose table type: [List format]
- Select a variable: REGION$
- Click OK

Count	Cum Count	Pct	Cum Pct	REGION$
9	21	18.0	42.0	Northeast
12	12	24.0	24.0	Midwest
16	37	32.0	74.0	South
13	50	26.0	100.0	West

© 1989, SYSTAT, Inc.

11.3 Confidence intervals

If your data are binomially or multinomially distributed, you may want confidence intervals on the cell proportions. The [Confi] option produces large sample confidence intervals on the cell proportions with the $1-\alpha$ level you specify. These confidence intervals are based on an approximation by Bailey (1980). SYSTAT uses that reference's approximation number 6 with a continuity correction, which fits closely the real intervals for the binomial on even small samples and performs well when population proportions are near 0 or 1. The confidence intervals are scaled on a percent scale for compatibility with the other **Tables** output.

Here is an example on data from Davis (1952) on the number of buses failing after driving a given distance (one of 10 distances). We print the percents of the 161 buses failing in each distance category to see the cover of the intervals. The filename is BUSES.

- Select **Tables/Tabulate...** from the **Stats** menu
- Choose table type: [Percent]
- Select a variable: DISTANCE
- Set [Confidence] to .95
- Click OK

```
Table of Values for DISTANCE

Percents of Total of this (Sub)Table
```

1	2	3	4	5	6
3.14	5.76	8.38	13.09	17.80	24.08

7	8	9	10	Total
17.28	8.38	1.05	1.05	100.00

```
Table of Values for DISTANCE

95 Percent Approximate Confidence Intervals Scaled as Cell Percents
```

1	2	3	4	5	6
8.23	11.87	15.26	21.00	26.45	33.42
.55	1.90	3.55	6.91	10.56	15.74

7	8	9	10
25.85	15.26	4.91	4.91
10.14	3.55	0.00	0.00

© 1989, SYSTAT, Inc.

11.4 Two-way table with tests and association measures

The DANISH data are from Svalastoga (1959), reprinted in Bishop, Fienberg, and Holland (1975). The variables FATHER and SON represent social class indices of fathers and sons, respectively. The variable COUNT gives the number of cases in each FATHER/SON combination.

FATHER	SON	COUNT
1	1	18
1	2	17
1	3	16
1	4	4
1	5	2
2	1	24
2	2	105
2	3	109
2	4	59
2	5	21
3	1	23
3	2	84
3	3	289
3	4	217
3	5	95
4	1	8
4	2	49
4	3	175
4	4	348
4	5	198
5	1	6
5	2	8
5	3	69
5	4	201
5	5	246

We weight the analysis by COUNT with **Data/Weight...**. This allows us to input the summary table without having to repeat the cell indices for every case. In our example, we have input the cells in a particular order. This is not necessary. In fact, we can repeat cells and counts in the input data. **Tables** totals all counts for a particular cell.

- Select [Show extended results] from **Data/Formats...** to request significance tests and measures of association
- Select **Weight** from the **Data** menu
- Select weighting variable: COUNT
- Click OK

© 1989, SYSTAT, Inc.

- Select **Formats...** from the **Data** menu
- Choose 0 decimal places to be shown
- Click OK

- Select **Tables/Tabulate...** from the **Stats** menu
- Select a row variable from the left selection list: FATHER
- Select a column variable from the right selection list: SON
- Click OK

The weighting specifies that each cell is to be weighted with the frequency given by COUNT. Incidentally, if COUNT has any non-integer values, they are truncated to integers before being used as weights for cell frequencies.

```
Table of  FATHER    (rows) BY     SON     (columns)

Frequencies

                1        2        3        4        5    TOTAL
     1         18       17       16        4        2       57
     2         24      105      109       59       21      318
     3         23       84      289      217       95      708
     4          8       49      175      348      198      778
     5          6        8       69      201      246      530
TOTAL          79      263      658      829      562     2391

Test statistic                    Value          DF         Prob
  Pearson Chi-square            754.104          16         .000
  Likelihood ratio chi-square   654.207          16         .000
  McNemar Symmetry Chi-square    24.421          10         .007

Coefficient                       Value    Asymptotic std error
  Phi                             .5616
  Cramer V                        .2808
  Contingency                     .4897
  Goodman-Kruskal Gamma           .5364                  .01852
  Kendall Tau-B                   .4057                  .01501
  Stuart Tau-C                    .3739                  .01425
  Cohen Kappa                     .2153                  .01375
  Spearman Rho                    .4701                  .01686
  Somers D     (Column dependent) .4048                  .01500
  Lambda       (Column dependent) .1159                  .02035
  Uncertainty (Column dependent)  .0965                  .00718
```

© 1989, SYSTAT, Inc.

11.5 Percents

Here is an example of percentage tables using the DANISH data.

- Select **Tables/Tabulate...** from the **Stats** menu
- Choose the types of tables to be produced: [Percent], [Row], and [Column]
- Select a row variable: FATHER
- Select a column variable: SON
- Click OK

Table of FATHER (Rows) by SON (Columns)

Percents of Total of this (Sub)Table

	1.	2.	3.	4.	5.	Total	N
1.	.75	.71	.67	.17	.08	2.38	57.00
2.	1.00	4.39	4.56	2.47	.88	13.30	318.00
3.	.96	3.51	12.09	9.08	3.97	29.61	708.00
4.	.33	2.05	7.32	14.55	8.28	32.54	778.00
5.	.25	.33	2.89	8.41	10.29	22.17	530.00
Total	3.30	11.00	27.52	34.67	23.50	100.00	
N	79	263	658	829	562	2391	

Table of FATHER (Rows) by SON (Columns)

Row Percents

	1.	2.	3.	4.	5.	Total	N
1.	31.58	29.82	28.07	7.02	3.51	100.00	57.00
2.	7.55	33.02	34.28	18.55	6.60	100.00	318.00
3.	3.25	11.86	40.82	30.65	13.42	100.00	708.00
4.	1.03	6.30	22.49	44.73	25.45	100.00	778.00
5.	1.13	1.51	13.02	37.92	46.42	100.00	530.00
Total	3.30	11.00	27.52	34.67	23.50	100.00	
N	79	263	658	829	562	2391	

© 1989, SYSTAT, Inc.

Table of FATHER (Rows) by SON (Columns)

Column Percents

	1.	2.	3.	4.	5.	Total	N
1.	22.78	6.46	2.43	.48	.36	2.38	57.00
2.	30.38	39.92	16.57	7.12	3.74	13.30	318.00
3.	29.11	31.94	43.92	26.18	16.90	29.61	708.00
4.	10.13	18.63	26.60	41.98	35.23	32.54	778.00
5.	7.59	3.04	10.49	24.25	43.77	22.17	530.00
Total	100.00	100.00	100.00	100.00	100.00	100.00	
N	79	263	658	829	562	2391	

11.6 Mantel-Haenszel test

If you select [Show extended results] in **Data/Formats...**, SYSTAT produces the Mantel-Haenszel statistic without continuity correction for any *k*×2×2 table. This tests the association between two binary variables controlling for a stratification variable. The stratification variable must be the first variable selected. The Mantel-Haenszel test is often used to test the effectiveness of a treatment on an outcome, to test the degree of association between the presence or absence of a risk factor and the occurrence of a disease, or to compare two survival distributions.

A study by Ansfield *et al.* (1977) examined the responses of two different groups of patients (colon or rectum cancer and breast cancer) to two different treatments.

CANCER$	TREAT$	RESPONSE$	NUMBER
"Colon-Rectum"	A	Positive	16.000
"Colon-Rectum"	B	Positive	7.000
"Colon-Rectum"	A	Negative	32.000
"Colon-Rectum"	B	Negative	45.000
"Breast"	A	Positive	14.000
"Breast"	B	Positive	9.000
"Breast"	A	Negative	28.000
"Breast"	B	Negative	29.000

© 1989, SYSTAT, Inc.

- Select **Weight...** from the **Data** menu
- Select weight variable: NUMBER
- Click OK

- Select **Tables/Tabulate...** from the **Stats** menu
- Select the stratifying variable from the left selection list: CANCER$
- Select row and column variables from the right selection list: TREAT$, RESPONSE$
- Click OK

```
Table of    TREAT$    (Rows) by RESPONSE$    (Columns)
   For the Following Values:
           CANCER$    = Breast

Frequencies

          Negative  Positive     Total
A               28        14        42
B               29         9        38
Total           57        23        80

Table of    TREAT$    (Rows) by RESPONSE$    (Columns)
   For the Following Values:
           CANCER$    = Colon-Rectum

Frequencies

          Negative  Positive     Total
A               32        16        48
B               45         7        52
Total           77        23       100

Mantel-Haenszel Chi-Square =      5.510     Probability = .019
```

The chi-square indicates a relation between treatment and outcome when effects of form of cancer are removed.

© 1989, SYSTAT, Inc.

11.7 Log-linear model

Let's model the father/son social class data from Example 11.4. We will fit a model that omits the father/son interaction. Be sure to tabulate these data with **Tables/Tabulate...** before attempting to run the following:

- Select **Tables/Log-linear models...** from the **Stats** menu
- Select variable: FATHER
- Click [+]
- Select variable: SON
- Choose [Fitted] and [Residuals] table types
- Click OK

```
Fitted Values

          1         2         3         4         5
1      1.88      6.27     15.69     19.76     13.40
2     10.51     34.98     87.51    110.26     74.75
3     23.39     77.88    194.84    245.48    166.41
4     25.71     85.58    214.10    269.75    182.87
5     17.51     58.30    145.86    183.76    124.58

Table of   FATHER     (Rows) by      SON      (Columns)

Standardized Residuals: (Observed-Fitted)/Sqrt(Fitted)

          1         2         3         4         5
1     11.74      4.29       .08     -3.55     -3.11
2      4.16     11.84      2.30     -4.88     -6.22
3      -.08       .69      6.75     -1.82     -5.54
4     -3.49     -3.95     -2.67      4.76      1.12
5     -2.75     -6.59     -6.36      1.27     10.88

Test of Fit of Model

   Degrees of Freedom =  16
   Pearson Chi-Square =      754.10    Probability =  .000
   Likelihood Ratio Chi-Square =     654.21    Probability =  .000
   Index of Dissimilarity=  19.427
```

Notice that the Pearson chi-square statistic is the same as that in Example 11.4 because we use an additive model, which amounts to a test for interaction.

© 1989, SYSTAT, Inc.

The index of dissimilarity is the percentage of cases that would have to be relocated to make the observed and expected counts equal. It is computed as 100 times the sum of the absolute differences, divided by twice the number of cases.

11.8 Log-linear model of multi-way tables

The following example demonstrates the use of **Log-linear models...** with a four-way table containing some empty cells. The data are from Morrison *et al.* (1973), cited in Bishop, Fienberg, and Holland (1975). The study involved breast cancer patients at three diagnostic centers (CENTER$), three age groups (AGE), whether they survived beyond three years after diagnosis (SURVIVE$), and the inflammation type (Minimum/Maximum) and appearance of the tumor (Malignant/Benign) combined under the variable INFLAPP$.

INFLAPP$	SURVIVE$	AGE	CENTER$	NUMBER
MinMalig	Dead	50	Tokyo	9
MinBengn	Dead	50	Tokyo	7
MaxMalig	Dead	50	Tokyo	4
MaxBengn	Dead	50	Tokyo	3
MinMalig	Alive	50	Tokyo	26
MinBengn	Alive	50	Tokyo	68
MaxMalig	Alive	50	Tokyo	25
.	.	.	.	.
.	.	.	.	.
MinMalig	Alive	70	Glamorgn	12
MinBengn	Alive	70	Glamorgn	11
MaxMalig	Alive	70	Glamorgn	4
MaxBengn	Alive	70	Glamorgn	1

First we tabulate the data:

- Select **Weight...** from the **Data** menu
- Select the weight variable: NUMBER
- Click OK

- Select **Tables/Tabulate...** from the **Stats** menu
- Select the first stratifying variable in the left selection list: CENTER$
- Select additional grouping variables in the right selection list: AGE, SURVIVE$, INFLAPP$
- Click OK

© 1989, SYSTAT, Inc.


```
Table of  SURVIVE$    (Rows) by  INFLAPP$    (Columns)
   For the Following Values:
            AGE     =       50
         CENTER$    = Boston

Frequencies

          MaxBengn  MaxMalig  MinBengn  MinMalig     Total
          --------  --------  --------  --------  --------
Alive            0         4        24        11        39
Dead             0         6         7         6        19
          --------  --------  --------  --------  --------
Total            0        10        31        17        58

Table of  SURVIVE$    (Rows) by  INFLAPP$    (Columns)
   For the Following Values:
            AGE     =       60
         CENTER$    = Boston

Frequencies

          MaxBengn  MaxMalig  MinBengn  MinMalig     Total
          --------  --------  --------  --------  --------
Alive            3        10        58        18        89
Dead             2         3        20         8        33
          --------  --------  --------  --------  --------
Total            5        13        78        26       122

Table of  SURVIVE$    (Rows) by  INFLAPP$    (Columns)
   For the Following Values:
            AGE     =       70
         CENTER$    = Boston

Frequencies

          MaxBengn  MaxMalig  MinBengn  MinMalig     Total
          --------  --------  --------  --------  --------
Alive            1         1        26        15        43
Dead             0         3        18         9        30
          --------  --------  --------  --------  --------
Total            1         4        44        24        73

Table of  SURVIVE$    (Rows) by  INFLAPP$    (Columns)
   For the Following Values:
            AGE     =       50
         CENTER$    = Glamorgn

Frequencies

          MaxBengn  MaxMalig  MinBengn  MinMalig     Total
          --------  --------  --------  --------  --------
Alive            1         8        20        16        45
Dead             0         3         7        16        26
          --------  --------  --------  --------  --------
Total            1        11        27        32        71
```

© 1989, SYSTAT, Inc.

```
Table of  SURVIVE$    (Rows) by  INFLAPP$    (Columns)
   For the Following Values:
            AGE      =       60
         CENTER$     = Glamorgn

Frequencies

        MaxBengn  MaxMalig  MinBengn  MinMalig     Total
Alive          4        10        39        27        80
Dead           0         3        12        14        29
Total          4        13        51        41       109

Table of  SURVIVE$    (Rows) by  INFLAPP$    (Columns)
   For the Following Values:
            AGE      =       70
         CENTER$     = Glamorgn

Frequencies

        MaxBengn  MaxMalig  MinBengn  MinMalig     Total
Alive          1         4        11        12        28
Dead           0         3         7         3        13
Total          1         7        18        15        41

Table of  SURVIVE$    (Rows) by  INFLAPP$    (Columns)
   For the Following Values:
            AGE      =       50
         CENTER$     = Tokyo

Frequencies

        MaxBengn  MaxMalig  MinBengn  MinMalig     Total
Alive          9        25        68        26       128
Dead           3         4         7         9        23
Total         12        29        75        35       151

Table of  SURVIVE$    (Rows) by  INFLAPP$    (Columns)
   For the Following Values:
            AGE      =       60
         CENTER$     = Tokyo

Frequencies

        MaxBengn  MaxMalig  MinBengn  MinMalig     Total
Alive          5        18        46        20        89
Dead           2        11         9         9        31
Total          7        29        55        29       120
```


© 1989, SYSTAT, Inc.


```
Table of  SURVIVE$    (Rows) by  INFLAPP$    (Columns)
   For the Following Values:
            AGE      =       70
         CENTER$     = Tokyo

Frequencies

        MaxBengn  MaxMalig  MinBengn  MinMalig     Total
Alive          1         5         6         1        13
Dead           0         1         3         2         6
Total          1         6         9         3        19
```

Now, let's fit a model. To avoid any cell having an expected value of zero, we add [Delta=.5] to each cell. The model involves main effects and three two-way interactions. We do not need to enter the main effects for the model as long as they appear in one or more of the interaction terms.

- Select **Tables/Log-linear models...** from the **Stats** menu
- Select variables from the selection list, clicking the [•] button between variables in the interactions, and the [+] button between interaction terms to build the following model: CONSTANT + AGE•CENTER$ + INFLAPP$•CENTER$ + SURVIV$•CENTER$
- Choose [Fitted] table type
- Specify Delta=.5
- Click OK

```
Model was fit after  2 iterations.
Table of  SURVIVE$    (Rows) by  INFLAPP$    (Columns)
   For the Following Values:
            AGE      =       50
         CENTER$     = Boston

Fitted Values

        MaxBengn  MaxMalig  MinBengn  MinMalig
Alive       1.41      4.69     24.38     10.94
Dead         .70      2.33     12.12      5.44

Table of  SURVIVE$    (Rows) by  INFLAPP$    (Columns)
   For the Following Values:
            AGE      =       60
         CENTER$     = Boston

Fitted Values

        MaxBengn  MaxMalig  MinBengn  MinMalig
Alive       2.86      9.53     49.54     22.23
Dead        1.42      4.74     24.63     11.05
```

© 1989, SYSTAT, Inc.

Table of SURVIVE$ (Rows) by INFLAPP$ (Columns)
For the Following Values:
AGE = 70
CENTER$ = Boston

Fitted Values

	MaxBengn	MaxMalig	MinBengn	MinMalig
Alive	1.75	5.82	30.28	13.59
Dead	.87	2.89	15.05	6.75

Table of SURVIVE$ (Rows) by INFLAPP$ (Columns)
For the Following Values:
AGE = 50
CENTER$ = Glamorgn

Fitted Values

	MaxBengn	MaxMalig	MinBengn	MinMalig
Alive	1.98	7.47	21.75	19.99
Dead	.92	3.48	10.12	9.30

Table of SURVIVE$ (Rows) by INFLAPP$ (Columns)
For the Following Values:
AGE = 60
CENTER$ = Glamorgn

Fitted Values

	MaxBengn	MaxMalig	MinBengn	MinMalig
Alive	2.98	11.25	32.76	30.12
Dead	1.39	5.24	15.25	14.02

Table of SURVIVE$ (Rows) by INFLAPP$ (Columns)
For the Following Values:
AGE = 70
CENTER$ = Glamorgn

Fitted Values

	MaxBengn	MaxMalig	MinBengn	MinMalig
Alive	1.19	4.48	13.05	11.99
Dead	.55	2.09	6.07	5.58

© 1989, SYSTAT, Inc.

Table of SURVIVE$ (Rows) by INFLAPP$ (Columns)
For the Following Values:
AGE = 50
CENTER$ = Tokyo

Fitted Values

	MaxBengn	MaxMalig	MinBengn	MinMalig
Alive	9.22	26.87	56.95	28.08
Dead	2.58	7.52	15.93	7.85

Table of SURVIVE$ (Rows) by INFLAPP$ (Columns)
For the Following Values:
AGE = 60
CENTER$ = Tokyo

Fitted Values

	MaxBengn	MaxMalig	MinBengn	MinMalig
Alive	7.38	21.50	45.56	22.46
Dead	2.06	6.01	12.74	6.28

Table of SURVIVE$ (Rows) by INFLAPP$ (Columns)
For the Following Values:
AGE = 70
CENTER$ = Tokyo

Fitted Values

	MaxBengn	MaxMalig	MinBengn	MinMalig
Alive	1.37	3.99	8.45	4.17
Dead	.38	1.12	2.36	1.17

Warning: More than one-fifth of fitted cells are sparse (frequency < 5)
Following significance tests are suspect.

Test of fit of model

Degrees of Freedom = 51
Pearson Chi-Square = 57.53 Probability = .246
Likelihood Ratio Chi-Square = 55.83 Probability = .298

Notice how SYSTAT warns you when the assumptions of the test are suspect. The method we used for dealing with empty cells is not entirely satisfactory.

Making the zeros structural with the [Zero] option is not appropriate for the CANCER data unless we believe that some combinations are impossible.

© 1989, SYSTAT, Inc.

Appendix I: Command reference

Introduction 508
Why commands? 508
Submitting 509
Creating a Batch 509
Submitting a Batch 510
General strategy 510
Background 510
Modules 512
Organization of the reference 513
HOT and COLD commands 514
DATA module 515
Command syntax 516
Command reference 520
File menu 520
Data menu 522
Goodies menu 524
Editor menu 525
Commands without menu equivalents 525
Cluster 527
Corr 529
Factor 532
MDS 534
MGLH 536
Nonlin 541
Npar 543
Series 545
Stats 550
Tables 551

© 1989, SYSTAT, Inc.

Appendix I: Command reference

You may sidestep SYSTAT's point-and-click menu interface by using the optional command interface. Just select **Show command window** from the **Goodies** menu. The Command window appears, and your first prompt ">" awaits orders.

The quickest way, perhaps, to learn the command line interface is to open the Command window and continue using the menus. The Command window follows along, showing you all the command line equivalents to your pointing and clicking. After a while, you'll get the idea and start using commands naturally.

This appendix explains the command syntax and defines all the options and arguments. First, we list some reasons you might want to use the command interface, and we give instructions for editing and submitting batch files. The remaining sections are references: lists of every command with syntax instructions, organized in parallel with the chapters in the manual.

© 1989, SYSTAT, Inc.

Why commands?

Why would you want to use the command interface? Because if you're doing a lot of repetitive analyses, generating a lot of variables algebraically, or using ranges of variables frequently, command lines are probably faster. If you're going to open the **Plot** dialog box and pick the same options, scales, and ranges repeatedly with different variables, chances are good that you could get done sooner typing.

What if you forget the syntax? Pull down a menu, do it once with the mouse, and see what shows up in the Command window!

Commands are better for a number of tasks:

Reproducing results
Suppose you're doing work that someone else will want to repeat. Perhaps you're doing consulting work, and your client wants to see a command file so she can verify your results. Perhaps someone wants to elaborate on your work, and he wants to see how you did it.

It would be difficult to give someone a comprehensive list of point-and-click instructions. Instead, you can hand over a list of exact commands. Many government agencies, in fact, *require* that command files be submitted with reports that include computer-generated results.

If all you want is a record of your work, you don't even need to use commands. Just open the Command window before you start working, do your usual point-and-click routine, and save or print the contents of the Command window. If you publish your results and a reader disputes your analyses, you can recreate your work from start to finish with SYSTAT's commands. Otherwise, you cannot defend your work. In the private sector, auditors often require lists of command statements. See Weber (1982).

Batching
You can write batch files of commands to be executed in succession. In other words, you enter a series of commands into a file and then submit the file, leave for coffee, and come back to get your results. The batch file could include commands to save and print results. You can also edit the command file and resubmit it.

© 1989, SYSTAT, Inc.

Modifying previous work
You can edit the Command window after an interactive session and resubmit the commands as a batch.

Submitting

The basic strategy for using batches is first to type a batch file or edit a portion of the Command window and then to submit that file or the portion from the window as a batch.

Creating a batch

There are two basic ways to create a batch: either type a bunch of commands from scratch, or edit something already in the Command window. The variations, though, are endless. You can type commands from scratch in any word processor and save them as a text file, or you can cut or copy them to the Clipboard. You can edit in either the SYSTAT Analysis window or the Command window. You can cut and paste between the windows. You can save your editing work in a file, or copy it to the Clipboard, or cut it to the Clipboard.

All windows in SYSTAT, except the graphics View window, are little text processors. You can do anything you want in them. If you have Stats results, for example, in the Analysis window and you want to use them elsewhere, you can cut them, type some comments around them, select a chunk, and print it. Snip out the comments, paste in some results from the correlation from half an hour ago (the one you still have in the window, because you remembered to "Scroll analyses"), select a chunk, save it to a file, and then edit it a little more into a set of commands. Copy it to the Clipboard, submit the Clipboard, sit back, and watch.

The catch: you have to clean out the garbage. Commands must look just as they do if you type them at prompts. No extra words, no extraneous characters (except spaces, commas, and tabs), no old prompts. So, if you're resubmitting some commands in the Command window, you must edit out the prompts and any SYSTAT messages (like "Series is cleared"). Each command should be on a separate line.

© 1989, SYSTAT, Inc.

Quick and dirty:

- Type or edit something into form:
  One command per line
  No extra junk—no prompts, comments, or weird characters
  Proper command syntax
- Save it in a file or copy it to the Clipboard:
  Select the text you want
- Use **File/Save selected text as...** to put it in a file
  Use **Edit/Copy** or **Edit/Cut** to put it in the Clipboard
  Use a word processor to save it in a text file

Submitting a batch

Once you've created a batch and saved it in a file or the Clipboard, you can submit it. SYSTAT processes the commands as if you were doing everything the usual, interactive way.

To submit a batch file, enter the command SUBMIT "*filename*", or use the mouse.

- Select **Submit command file...** from the **File** menu
- Specify the file to be used
- Click Submit

To submit a batch from the Clipboard, type Command-F or use the mouse.

- Select **Submit Clipboard** from the **File** menu

General strategy

We discuss command syntax, rules for how to type commands, and guidelines for deciphering our notation in the section "Command syntax," below. Here, we discuss the general strategy for using SYSTAT's command interface.

Background

Historically, SYSTAT has been a modular, command-driven program. Different categories of operations, analyses, etc., were grouped into separate modules which were effectively separate programs. These modules were very closely related and could exchange information with each other smoothly, but the user still needed to switch from module to module when doing different types of tasks.

© 1989, SYSTAT, Inc.

SYSTAT has gotten much praise, in fact, for its compact and very efficient modular design. Among other advantages, the modular design allows users of Mac versions, or PC/compatible versions, or mainframe versions to switch comfortably from machine to machine. Aside from needing to make minor adaptations of data file formats, they can take the same data sets, command files, and manuals from machine to machine.

Until this version, even Macintosh SYSTAT has been mostly command-driven. This version makes one major change from all previous versions of SYSTAT for any platform in that it is completely integrated into one environment. You do statistics, tables, regression, graphs, and even data entry in one program. You never have to do anything more complicated than go to a different menu.

By now, you've no doubt seen some of the exciting advantages of such an integrated environment: the Data Editor and the graphics View window are linked in such a way that you can see the relationships between data cells and plot points *visually*, instantly. You can move effortlessly from analysis results—lines of numbers, words, *p*-values, and other dry information—to the graphs that *show* you their meanings.

However, because commands are more efficient for some kinds of tasks—like repetitive number-crunching, sophisticated analyses, and keeping records of your work, we still offer a command interface even for this point-and-click, integrated Macintosh version.

Perhaps the users who will make most use of the command interface are those who have used previous versions and those who still use command-driven versions of SYSTAT on other types of machines (alas, not all the world's computer users own a mouse!) for some of their work. For these users, lateral compatibility (compatibility across different machine and operating system platforms) is very important. Therefore, we have decided to make the command interface as similar as possible to the traditional modular design of SYSTAT.

This means that, even though Macintosh SYSTAT is one integrated program, our command interface still calls the old "modules." For example, the first command you have to issue to do basic statistics is STATS. STATS switches you to the (invisible) STATS module and tells SYSTAT that the next series of commands you are going to use will be for operations that would use the **Stats** item in the **Stats** menu.

© 1989, SYSTAT, Inc.

Modules

We have named all of the items in the **Stats** menu after the old modules, for two reasons. First, long-time SYSTAT users will have no trouble finding their way around the menus. Second, new and experienced users alike will easily remember the "module name" that they have to type before they use commands for something. Note that the **Stats** and **Graph** menus are, essentially, SYSTAT and SYGRAPH menus. The remaining menus are all global items that have always been available in SYSTAT and SYGRAPH alike.

The first step when using commands for anything is the same as the first step without commands: open a data file. The command equivalent for **File/Open...** is USE:

```
USE MEDICAL
```

SYSTAT remembers what file you opened with USE even when you switch modules. The files always follow you around through the various parts of SYSTAT and SYGRAPH, just as they do when you are using the usual mouse interface.

If you need to do anything to "adjust" the way you use data, you can do it at any time. So, if you need **Data/By groups...**, **Select...**, or **Weight...**, you can do that now with commands like the following:

```
WEIGHT COUNT
SELECT REGION$='New England'
BY GROUP1
```

You can start a SELECT, a BY, or a WEIGHT at any time. These three things stay in effect across modules. So, to turn them off, you have to either issue the command again, but without an argument, or you have to USE a new data file (or use the same one again).

```
WEIGHT     or     USE MEDICAL
SELECT
BY
```

You can also "turn off" a BY, SELECT, or WEIGHT command by issuing a new one:

© 1989, SYSTAT, Inc.


```
WEIGHT X(1)
SELECT GROUP=3
BY REGION$
```

When in doubt, start fresh: USE the file anew. That way, you're sure that you're using the whole dataset in your analyses.

The next thing you must do is "switch to a module." So, if you are about to do an analysis that ordinarily you would find under **Corr** in the **Stats** menu, your first command is:

```
CORR
```

Next, you type the command or commands needed to do the analysis. Suppose you want to get the Pearson correlation matrix of the three variables ACCIDENT, CARDIO, and CANCER. The command you need is the same as the menu item's name (**Corr/Pearson...**). The first thing you would do in the dialog box is select variables, and that's the first thing you do with the command.

```
PEARSON ACCIDENT CARDIO CANCER
```

You may need to use *options* with the command. *Options* are generally the miscellaneous items in a dialog box that have checkboxes, or radio buttons, or little boxes where you type numbers. Options for SYSTAT commands come after a slash (/). We'll discuss command syntax in more detail later. For now, let's just consider our PEARSON example. Suppose we would click "Pairwise" deletion in the dialog box. The command equivalent is obvious:

```
PEARSON ACCIDENT CARDIO CANCER / PAIRWISE
```

Organization of the reference

To help ease your transition to the command interface, we have organized the command reference to follow the organization of the menus and dialog boxes.

© 1989, SYSTAT, Inc.

In the left margin of each page, where you usually see section headings, we have shown the **Stats** menu equivalent. Commands are grouped according to "module," or menu item. We show the menu item, like **Corr** or **Nonlin** or **Tables**, in boldface. Within each boldface module/section heading are the menu items, in regular lightface type: things like Pearson..., Loss function..., or Tabulate... The command or commands that you use are shown grouped according to these subitems (dialog boxes).

You may sometimes need to skim the descriptions for all the commands within a subitem section to figure out which ones you need, but usually the names of the commands are similar enough to what you see in the dialog box that you shouldn't have any trouble.

Since **MGLH** is the most powerful and flexible part of SYSTAT, it is also the most complicated. We have organized the **MGLH** section of the command reference a little bit differently: we divide the commands into tasks: "Model estimation commands" and "Hypothesis testing commands."

Also, remember: you can always get help for any command. Just type "HELP *command*," where *command* is the name of the command that confuses you. For instance, if you need help with PEARSON or any of its options:

```
HELP PEARSON
```

Finally, if you are in doubt about how to use commands to accomplish a particular type of analysis, look in the manual for an example that is similar to what you are trying to do. Make a note of its number—e.g., Example 7.3 or 8.6, etc. The next appendix, **Command equivalents for examples**, shows the command equivalents to the mouse instructions we give for each example in the manual. The examples are grouped by chapter and retain the same numbering as in the chapters.

HOT and COLD commands

Some commands are HOT. Some commands are COLD. When you type a HOT command you see an immediate action—calculations are done, the screen fills with statistics or graphics, the printer starts working, etc. COLD commands turn switches on and off, specify conditions, and select options.

© 1989, SYSTAT, Inc.

Generally, you use COLD commands to set up a test or (in DATA) a data manipulation, and the last command you type is a HOT command that sets the procedure in motion. Since the COLD commands all just set up the procedure, you can usually put them in any order. The important thing is that the HOT command comes last.

For example, **MGLH** procedures usually involve several COLD commands followed by one HOT command. The COLD commands set up the model, and the HOT command (ESTIMATE) tells SYSTAT to start crunching numbers and giving you answers.

```
MGLH
USE ROTATE
MODEL RT = CONSTANT + ANGLE
SAVE RESIDS
ESTIMATE
```

MGLH switches you to the MGLH "module." USE ROTATE opens a datafile. (You could do USE before MGLH. It doesn't matter.) The MODEL equation sets up the model. SAVE requests that residuals be saved in a file named RESIDS. (You could do MODEL and SAVE in either order.) ESTIMATE tells SYSTAT to start working.

Again, when in doubt, use HELP and check the **Command equivalents for examples** appendix, next.

DATA module

Remember that the DATA facilities for powerful data manipulations are documented in the *DATA* volume, which has its own command reference, Appendix I of that volume.

DATA is also a "module." You can move into DATA just as you would move into MGLH or CORR or FACTOR.

© 1989, SYSTAT, Inc.

Command syntax

"Syntax" is grammar. Command syntax is all the typing rules: this word has to come first, slashes come later, quotation marks do this, filenames have to look that way, and so on.

Prompts

The "greater than" sign (>) in the Command window is your "prompt." A prompt tells you that SYSTAT is waiting for you to tell it something. After you type a command, press Return. If you typed a HOT command, SYSTAT begins work and you do not see another prompt until SYSTAT is done. If you typed a COLD command, you get another prompt right away.

Upper- and lower-case

You can type commands in upper or lower-case. Commands are not case sensitive except when they include text values, which *are* case sensitive.

Abbreviations

Command words can be abbreviated to the first three letters, but function names and variable names must always be typed fully. Commas, spaces, and tabs are interchangeable.

Order

Commands consist of several parts. The first is the command word: the part that says what the command is supposed to do. "PEARSON" is a command word that means "find Pearson correlation coefficients."

The next part is the "argument:" the object of the command. The argument answers the question, "Find the correlation coefficients for *what*?!" You answer the question by listing some variables:

```
PEARSON ACCIDENT, CARDIO, CANCER
```

Sometimes, arguments are optional; for instance, if you don't say which variables, PEARSON finds the coefficients for all the numeric variables it can find.

The last part of any command is the option list. The option list is (how could you guess?) optional.

© 1989, SYSTAT, Inc.

Put a slash before the option list and then list all the options you want. You only need one slash, no matter how many options you use. Each command has a set of options. Some options are exclusive—if you pick A, you can't pick B. These are listed together in our reference, separated with bars: LISTWISE | PAIRWISE. You can pick LISTWISE, or you can pick PAIRWISE, or you can skip it entirely and SYSTAT uses its *default*: the thing it does if you don't tell it otherwise. More on defaults below.

Placeholders

Placeholders tell you what to type. For instance, we show the placeholder *yvar* where you're supposed to type a dependent variable name. Placeholders are always shown in *italic* typeface.

depvar	dependent variable (vertical axis for graphs)
indvar	independent variable (horizontal axis for graphs)
var$	character variable
gvar	grouping variable
varlist	one or more variables (variations: *xvarlist*, *yvarlist*, *gvarlist*, *var$list*)
"*filename*"	any valid filename; must be surrounded by double or single quotation marks if it contains path (folder) information, spaces, or symbols
"*char*"	one or more characters, for instance a value for a character variable; must be surrounded by double or single quotation marks
exprn	any mathematically valid expression

Brackets

Items enclosed in brackets are optional. We don't bother putting brackets around the list of options, since it is understood that options are optional. Usually, things that involve brackets are explained in detail in the description of the command to the right.

Quotation marks

Quotation marks that are included in the syntax demonstrations and in example commands are ***required***. Quotation marks are required for all text values and filenames.

© 1989, SYSTAT, Inc.

Long commands
If a command is too long to fit on one line, put a comma at the end of the line, press Return, and continue typing. You can stretch a single command down several lines as long as you keep using commas.

Slashes
Slashes set off options, such as graph dimensions, symbol sizes, etc. You *must* type the slash as part of the command. Use only one slash per command, regardless of the number of options; in other words, the slash divides the command-argument half of the command from the option half.

Items that follow slashes are understood to be optional. For clarity, we do not enclose the lists of options inside brackets.

For example, here is part of the entry for STATISTICS.

STATISTICS *varlist*

/MEAN MINIMUM
MAXIMUM SD CV
KURTOSIS RANGE
SEM SKEWNESS
SUM VARIANCE

All of the following are valid commands. (Since SYSTAT only reads the first three letters of command names, we can abbreviate to STATS.) Note that commas and spaces are interchangeable.

```
STATS A,B
STATS A,B/MEAN, MIN,, KUR
STATS  /      MEAN MIN,KUR RANG
STATS/SUM,VAR
```

Vertical bars
A vertical line (|) means "or." For example, for many items in CORR, you can use listwise or pairwise deletion of missing values:

PAIRWISE | LISTWISE

© 1989, SYSTAT, Inc.

Options that are separated by vertical bars are exclusive. You can use only one at a time (or none). For example, TICK=IN | OUT means that you can have tick marks placed inside *or* outside the axes, but not both.

Defaults
Defaults are preset option selections that take effect unless you change them. Defaults are underlined. For example, in the option above, the underlining indicates that pairwise deletion of missing values is the default and will be done unless you specify LISTWISE.

Sometimes you see underlining without the vertical "or" bar. This indicates a default which doesn't necessarily exclude other options listed with it. For example, the reference entry for STATISTICS (equivalent to **Stats/Statistics...** in the **Stats** menu) begins like this:

STATISTICS *varlist*

/MEAN MINIMUM
MAXIMUM SD CV
KURTOSIS RANGE
SEM SKEWNESS
SUM VARIANCE

MEAN, MINIMUM, MAXIMUM, and SD are underlined because if you don't include any options in your command, you will get mean, minimum, maximum, and standard deviation statistics. If you include any options, you will get only that statistic. For example, the following would produce range and skewness statistics *only*:

```
STATISTICS ACCIDENT CARDIO CANCER/RANGE SKEW
```


© 1989, SYSTAT, Inc.

File menu Open...	**EDIT** "*filename*"	Opens data file *filename* and displays it in Data Editor. If you do not specify *filename*, SYSTAT opens a new (empty) worksheet. HOT.
	GET "*filename*"	Opens text file *filename*. Use in conjunction with INPUT command. COLD.
	USE "*file1* " [(*varlist*)]	Opens a data file without the Data Editor. If you include *varlist*, only those variables are opened. You must enclose *varlist* in parentheses. HOT.
Import...	**IMPORT** *filename* [(*varlist*)]	Translates *filename* into SYSTAT format. You can optionally include *varlist* if you want to import only certain variables from the file.
	/ TEXT | MAP | EXCEL	Specifies importing format. TEXT, the default, reads plain text (ASCII) files. MAP reads map data into .MAP files (see Appendix V in the *SYGRAPH* volume for information about map files). EXCEL translates Microsoft Excel files into SYSTAT files.
Save as...	**MAC** "*filename*"	Saves current file into tab delimited text file named *filename*. HOT.
	PUT "*filename*"	Creates a text file of the current data file. The text file is comma delimited with quotation marks around character values. COLD.
	WRITE "*filename*"	Saves current worksheet in SYSTAT file named *filename*. HOT.

© 1989, SYSTAT, Inc.

Save graph as...	**PICT** "*filename*"	Saves graph currently in View window in PICT file named *filename*. HOT.
Submit command file...	**SUBMIT** "*filename* "	Reads and executes commands contained in *filename*. *Filename* must be a text file. HOT.
	/ECHO	Displays the commands in *filename* as they are processed.
Quit	**QUIT** [= * | @ | "*filename*"	Exits SYSTAT to the Finder. The optional arguments let you output the command log to a specific destination. Asterisk (*) sends the command log to the screen only. At-sign (@) sends the command log to a printer and the screen. "*Filename*" sends the command log to the screen and a text file called *filename*. Not available in the Data Editor. HOT.

© 1989, SYSTAT, Inc.

Data menu Rank...	**RANK** *varlist*	Transforms values of all variables in *varlist* to their ranked values. Use "SAVE *filename*" before RANK to specify the output file. If you omit *varlist*, all variables are ranked. HOT.
Sort...	**SORT** *varlist*	Sorts file by *varlist*. If you include more than one variable, SYSTAT does a nested sort using variables in the order you specified. Use "SAVE *filename*" before SORT to specify the output file. You may use numeric and character variables. HOT.
Standardize...	**STANDARDIZE** *varlist*	Standardizes all variables in *varlist*. Specify an output file using "SAVE *filename*" before STANDARDIZE. If you omit *varlist*, all variables are standardized. HOT.
Transpose...	**TRANSPOSE**	Transposes the current file. The file can contain up to 99 variables and must contain only numeric variables.Use "SAVE *filename*" before TRANSPOSE to specify the output file. HOT.
By groups...	BY *gvarlist*	Specifies grouping variables for subgroup analysis. Analyses are performed separately for each subgroup. BY is not available in the Data Editor, MDS, or SERIES. COLD.
	/MISS	Excludes from analysis any case with a missing value for a BY variable.
	PAGE	Begins output on a new page for each subgroup.

© 1989, SYSTAT, Inc.

Select...	**SELECT** *exprn1* [...]	Selects a subgroup of cases for analysis. You may include up to 10 expressions. Only cases meeting all the expression conditions are used. Using the SELECT command without an argument ends current selection conditions. COLD.
Weight...	**WEIGHT**=*var*	Weights each observation with the value of *var*. The sample size for analyses is the sum of the values of *var*. The values of *var* are truncated to integers before processing. Issuing the WEIGHT command without an argument clears the current weight selection. Not available with the MDS or SERIES items. COLD.
Results to screen/ printer/file as...	**OUTPUT**=* | @ | "*filename*"	Sends subsequent output to the specified device. Asterisk (*), the default, sends output to the screen only. At-sign (@) sends output to a printer and the screen. *Filename* sends output to the screen and a text file called *filename*. Not available in the Data Editor. COLD.
Notes...	**NOTE**=*n* | '*line1*' ['*line2* ...]	Prints any note (character string) in the text output, or the printer or file if output is redirected. HOT. NOTE can print ASCII characters by their index; specify the index number without quotation marks. For instance, NOTE=13 puts a carraige return in the output. (You can specify both an index and a character string in a single line.) HOT.

© 1989, SYSTAT, Inc.

Page formats...	**PAGE**	Selects screen display and printer output format and characteristics with the options that follow. COLD.
	/FILE=*n*	Specifies the number of lines per page in a text file. The default value is 56. Specify 0 for no page breaks.
	PRINTER=*n*	Specifies the number of lines per printed page. The default value is 56. Specify 0 for no page breaks.
	NARROW I WIDE	Specifies the number of columns used for analysis (text) output. NARROW, the default, specifies 80 columns; WIDE specifies 132.
	TITLE='*line1*', '*line2*', ...	Specifies a title for each page of output. You may specify up to 10 title lines; each line may be up to 132 characters long. SYSTAT centers each line on the page.
Formats...	**FORMAT=*n***	Specifies the number of decimals to display in output ($0 \le n \le 9$). The default is 3. COLD.
	/UNDERFLOW	Prints tiny numbers that otherwise would appear as "0" in exponential notation.
Goodies menu Show options in effect	**OPTIONS**	Displays options currently in effect. Not available in the Data Editor. HOT.

© 1989, SYSTAT, Inc.

Editor menu Math...	**LET** *var=exprn*	Assigns value specified by *exprn* to *var*. HOT.
Recode...	**IF** *exprn* **THEN LET** *var=exprn*	If *exprn* is true, the LET statement is executed. HOT.
Find case... Find next	**FIND** *exprnlist*	Finds the first case below the cursor's position that meets the condition specified by *exprnlist*. If you do not include *exprnlist*, SYSTAT uses the previous *exprnlist*. HOT.
Fill worksheet...	**REPEAT** *n*	Fills *n* cases in the worksheet with missing values. You must label variables before you use REPEAT. HOT.
Delete...	**DELETE**	Prevents current case from being saved into a new file. Usually used with IF-THEN; e.g., IF X>10 THEN DELETE. COLD.
Formats...	**FORMAT=*n***	Specifies the number of decimals to display in the data editor. The default is 3. HOT.
	TYPE=CORRELATION I COVARIANCE I DISSIMILARITY I RECT I SIMILARITY I SSCP	Specifies the type of matrix being entered. All but RECT are triangular matrices such as are saved by procedures in CORR.
Commands without menu equivalents	**HELP** *command*	Displays help for specified *command*. HOT.
	NAMES	Displays variable names in current file. Not available in the Data Editor. HOT.

© 1989, SYSTAT, Inc.

REM *comment*	REM lets you put comments in command files. All text on the same line is ignored.
RSEED=*n*	Specifies the random number seed *n* to be used for any random data generated by the SYSTAT functions (used with IF...THEN and LET). You can specify any integer between 1 and 30,000.

© 1989, SYSTAT, Inc.

Cluster	**LABEL**=*var$*	Labels each case in the JOIN tree or KMEANS output with the value of *var$*. COLD.
	PRINT=LONG | SHORT	Enter PRINT=LONG before JOIN command to print cluster joining distances in addition to standard output. COLD.
	SAVE "*filename*"	Saves cluster membership identifiers for JOIN (by rows) or KMEANS. All standard options, plus the following, apply. COLD.
	NUMBER=*n*	Specifies the number of cluster identifiers to save. Required with JOIN. Has no effect with KMEANS.
	SINGLE | DOUBLE	Specifies whether to save in single or double precision.
Join...	**DISTANCE**=EUCLIDIAN | GAMMA | PCT | PEARSON	Specifies the distance metric for hierarchical clusters (*see* JOIN). EUCLIDEAN is the default and specifies a normalized Euclidian distance metric. GAMMA specifies a distance metric of $1-\gamma_{ij}$ (where γ_{ij} is the Goodman-Kruskal gamma coefficient). PCT specifies a distance metric which is the percentage of comparisons that result in disagreements. PEARSON specifies a distance metric $1-r_{ij}$ (where r_{ij} is the Pearson correlation). COLD.

© 1989, SYSTAT, Inc.

	LINKAGE=AVERAGE | CENTROID | COMPLETE | MEDIAN | SINGLE | WARD	Specifies the method of calculating cluster distances for the JOIN command. AVERAGE averages all distances between pairs of objects in different clusters. CENTROID uses the average value of all objects in a cluster (the cluster centroid) as the reference point for distances to other clusters. COMPLETE uses the most distant pair of objects in two clusters to compute between-cluster distances. MEDIAN computes distances by median linkage method. SINGLE uses the closest pair of objects in two clusters to compute between-cluster distances. WARD computes distances by Ward's method. COLD.
	JOIN *varlist*	Produces hierarchical clusters. HOT.
	/COLUMNS | ROWS | MATRIX	COLUMNS clusters columns of data. ROWS clusters rows of data. MATRIX clusters rows and columns at the same time.
	SYMBOL=*n*	Specifies number of symbols used in output. Available only with MATRIX option.
Kmeans...	**KMEANS** *varlist*	Produces K-means clustering. HOT.
	/ITER=*n*	Specifies the number of iterations used to assign cases to groups. The default is 50.
	NUMBER=*n*	Specifies the number of clusters. Must be a positive integer equal to or less than the number of cases. The default is 2.

© 1989, SYSTAT, Inc.

Corr	**SAVE** "*filename*"	Saves the next matrix that you compute. All standard SAVE options apply. COLD.
	SINGLE | DOUBLE	Specifies whether to save in single or double precision.
Clty...	**CITY** *varlist*	Computes a matrix of city block distances for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.
Covariance...	**COVARIANCE** *varlist*	Computes a covariance matrix for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.
Dichotomous...	**S2** *varlist*	Computes a positive matching dichotomous coefficient matrix for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.
	S3 *varlist*	Computes Jaccard's dichotomous coefficient matrix for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.
	S4 *varlist*	Computes simple matching dichotomous coefficient matrix for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.

© 1989, SYSTAT, Inc.

	S5 *varlist*	Computes Anderberg's dichotomous coefficient matrix for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.
	S6 *varlist*	Computes Tanimoto's dichotomous coefficient matrix for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.
Euclidean...	**EUCLIDIAN** *varlist*	Computes a normalized Euclidian distance (root-mean-squared differences) matrix for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.
Gamma...	**GAMMA** *varlist*	Computes a Goodman-Kruskal gamma coefficients matrix for ordered categories for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.
Mu2...	**MU2** *varlist*	Computes a Guttman mu2 matrix of monotonicity coefficients for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.

© 1989, SYSTAT, Inc.

Pearson...	**PEARSON** *varlist*	Computes a Pearson correlation matrix for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.
	BPROB | PROB	BPROB prints a Bonferonni adjusted probability for each correlation and the Bartlett chi-square statistic for an identity matrix. PROB prints a probability for each correlation and the Bartlett chi-square statistic for an identity matrix.
Spearman...	**SPEARMAN** *varlist*	Computes a Spearman rank order correlation matrix for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.
SSCP...	**SSCP** *varlist*	Computes a sum of squares and cross products matrix for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.
Tau...	**TAU** *varlist*	Computes a Kendall tau-b matrix for all variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.
	/LISTWISE | PAIRWISE	Specifies how missing values are deleted.

© 1989, SYSTAT, Inc.

Factor	**PRINT**=LONG ! <u>SHORT</u>	LONG prints the matrix to be factored in addition to the standard output. SHORT, the default, does not print the matrix with the output. COLD.
	SAVE "*filename*"	Saves specified statistic into *filename*. You cannot save when USEing a correlation or covariance matrix. All standard SAVE options apply. COLD.
	/COEF ! LOADINGS ! <u>SCORES</u>	COEF saves component score coefficients. LOADINGS saves component or factor loadings. SCORES, the default, saves component scores; it is not available with the common factor model.
	SINGLE ! <u>DOUBLE</u>	Specifies whether to save in single or double precision.
	TYPE=<u>CORRELATION</u> ! COVARIANCE	Specifies the type of matrix to factor. Correlation is the default. COLD.
	/<u>LISTWISE</u> ! PAIRWISE	Specifies how missing values are deleted for matrix computation. Listwise is the default.
Principal components...	**EIGEN**=*n*	Specifies the smallest eigenvalue to retain. Only components with eigenvalues larger than *n* are printed and rotated. The default is 0. If both EIGEN and NUMBER are used, the restriction resulting in the smaller number of components is used. COLD.
	NUMBER=*n*	Specifies the number of factors to print and rotate. *N* must be an integer between 1 and the number of variables specified in the FACTOR command, inclusive. The default value is the number of variables. If both EIGEN and NUMBER are used, the restriction resulting in the smaller number of components is used. COLD.

© 1989, SYSTAT, Inc.

	ROTATE=EQUAMAX ! QUARTIMAX ! <u>VARIMAX</u>	Rotates factors by the specified method; the default is no rotation, but if rotation is specified without an argument, VARIMAX is used. COLD.
	FACTOR *varlist*	Computes factor calculation (principal components analysis) on variables in *varlist*. All numeric variables are used by default. HOT.
(Common factor model...)	/IPA	Estimates the common factor model with iterated principal axis factoring.
	ITER=*n*	Specifies the maximum number of iterations in converging on the solution. The default is 25.
	TOL=*n*	Specifies the convergence criterion *n* where $0 \leq n \leq 1$. The default value is .001. If both ITER and TOL are specified, the limit that results in fewer iterations is used.
	PLOT	Produces a scree plot and factor loadings plots (factors plotted against original variables).
	LABEL	Labels the points in the loadings plots with the original variable names.
	SORT	Sorts the rows of the loadings matrix so that loadings greater than .5 on each component are adjacent.

© 1989, SYSTAT, Inc.

MDS	**SAVE** "*filename*"	Saves the matrix of interpoint distances or the final configuration. You must enter SAVE before SCALE. All standard SAVE options apply. COLD.
	/CONFIG	Saves the final configuration.
	DIST	Saves the interpoint distance matrix.
	RESID	Saves observed and fitted distances.
	SINGLE | DOUBLE	Specifies whether to save in single or double precision.
Configuration...	**CONFIGURATION** *matrix*	Specify a starting configuration for the scaling. There must be as many rows as items and columns as dimensions. You must use the DIMENSION command before CONFIGURATION. COLD.
Scale...	**DIMENSION**=*n*	Specifies the number of dimensions in which to scale. The default is 2. *N* must be a positive integer less than or equal to the number of variables that you scale. COLD.
	DECREMENT=*n*	Sets the minimum decrement in stress or alienation at which iterations terminate. The default is zero. COLD.
	ITERATIONS=*n*	Specifies the maximum number of iterations allowed before termination. The default is 50. COLD.
	METHOD=GUTTMAN | KRUSKAL	Selects the method of scaling. GUTTMAN calculates the Guttman Coefficient of Alienation. KRUSKAL, the default, uses Kruskal's stress formula 1. COLD.

© 1989, SYSTAT, Inc.

Appendix I: Command reference

MINIMUM=*n*	Specifies the stress or alienation value at which iterations terminate. The default is zero. COLD.
R=*n*	Specifies the constant for the Minkowski (power) metric for computing distances. The default is 2 (the Euclidian distance metric). A constant of 1 is the city block metric. COLD.
REGRESSION = MONOTONIC | LINEAR	Specifies the form of the function relating distances to similarities (or dissimilarities). Available only with the Kruskal method. MONOTONIC, the default, specifies nonmetric scaling, while LINEAR specifies metric scaling. COLD.
SHEPARD=1 | 2	Specifies the number of Shepard plots to print. The default is 1. With 2, the plots identify the pair of items that comprise each point. COLD.
SCALE *varlist*	Initiates scaling on specified variables. If you omit *varlist*, all variables are used. HOT.

© 1989, SYSTAT, Inc.

MGLH	**PRINT**=LONG | <u>SHORT</u>	LONG produces additional statistics depending on the model being estimated. SHORT prints the default statistics. COLD.
	SAVE "*filename*"	Saves specified statistics into *filename* and displays the Durbin-Watson statistic, the first-order autocorrelation, and any outliers. COLD.
	/ADJUST | COEF | MODEL | <u>RESID</u>	ADJUST saves adjusted cell means from analysis of covariance. COEF saves the regression coefficients. RESID, the default, saves residual diagnostics. For a univariate model (one dependent variable), this includes the estimate, residual, leverage, Cook's d, Studentized residual, and standard error of prediction for each case. For a multivariate model, SYSTAT saves the estimate, residual, and leverage. MODEL saves the variables from your MODEL statement in addition to the statistics given by RESID.
	SINGLE | <u>DOUBLE</u>	Specifies whether to save in single or double precision.
Model estimation commands	**ANOVA** *var*	Specifies *var* as a dependent variable in a fully-factorial analysis of variance. **Note:** CATEGORY, ANOVA, and ESTIMATE are used together, in order, to produce analysis of variance. Add the COVARIATE command between CATEGORY and ANOVA for analysis of covariance. COLD.
	REPEAT=*n,m*	Specifies the levels of the repeated measures factors.

© 1989, SYSTAT, Inc.

Appendix I: Command reference

CATEGORY *gvar1=n, gvar2=m, gvar3=p, ...*	Specifies the number of levels *n, m, p, ...*, for categorical variables *gvar1, gvar2, gvar3, ...*You must specify all independent categorical variables (factors) in a CATEGORY statement before MODEL (or ANOVA). Categorical variables must be numeric with values 1, 2, 3, ..., *k*, where *k* is the number of levels in the variable. Use value of zero only for nested terms in models. COLD.
COVARIATE *var*	Specifies *var* as a covariate in a fully-factorial analysis of variance. Available only with the ANOVA command. COLD.
MODEL *varlist1* = [CONSTANT] + *varlist2* + var1 * *var2, ...*	Specifies the linear model to estimate. *Varlist1* specifies the dependent variable(s). CONSTANT is an optional parameter (when in doubt, include the CONSTANT). *Varlist2* specifies main effects (independent variables). Specify interactions by linking variables with an asterisk. COLD.
/N=*n*	When your data file is a symmetric matrix (e.g., a covariance matrix), specify the sample size *n* that generated the matrix.
REPEAT=*m, n*	Specifies the number of levels for each repeated measures factor; *m* specifies the number of levels for the first factor and *n* for the second. You cannot specify more than two repeated measures factors.
ESTIMATE	Tells SYSTAT to estimate the MODEL or ANOVA you specified. HOT.
/MIX	Estimates a mixture model. Your independent variables must sum to a constant to use this.

© 1989, SYSTAT, Inc.

	STEP	Initiates stepwise variable selection. HOT.
	/ENTER=*n*	Includes any independent variable that meets or exceeds the alpha (*n*) level of significance. The default is 0.015. *N* must be between 0 and 1 and less than or equal to the value specified in REMOVE.
	FORCE=*n*	Forces the first *n* independent variables into the equation.
	REMOVE=*n*	Excludes any variables from the equation if the significance of its contribution falls below *n*. The default is 0.015. *N* must be between 0 and 1 and greater than or equal to the value specified in ENTER.
Hypothesis testing commands	**HYPOTHESIS**	Tells SYSTAT you want to test hypotheses on a previous MODEL. You must enter the HYPOTHESIS command before you can use any of the following commands. COLD.
	ALL	Tests all of the coefficients in the model.
	AMATRIX *matrix*	Specifies a design matrix for your hypothesis (for between subjects effects). COLD.
	CMATRIX *matrix*	Specifies a design matrix for the dependent variables. COLD.
	CONTRAST *matrix*	Specifies contrasts on a single effect. COLD.
	DMATRIX *matrix*	Identifies the value of the null hypothesis. The default value is 0 for each effect. COLD.

© 1989, SYSTAT, Inc.

EFFECT=*varlist, var1 & var2, ...*	Specifies the effect(s) to include in the hypothesis test. *Varlist* specifies main effects. Interactions are specified by joining two variables with the and sign (&). COLD.
ERROR=*varlist, var1 & var2, ...*	Specifies an error term for hypothesis testing. *Varlist* specifies main effects. Interactions are specified by joining two variables with the and sign (&). COLD.
FACTOR=HYPOTHESIS | ERROR	Identifies the matrix to factor. You must specify one of the options. HYPOTHESIS factors the sum of product hypothesis matrix. ERROR factors the residual correlation, covariance or SSCP matrix (depending on the type of file used). COLD.
PRINT=LONG | <u>SHORT</u>	SHORT, the default, includes significant tests only. LONG includes additional matrix output when you test hypotheses. This is provided automatically with discriminant analysis. COLD.
PRIORS=*m, n, p, ...*	Specifies prior probabilities of group membership in discriminant analysis. *M* corresponds to the first level of your grouping variable, *n* to the second, and so on. The default is equal probability for each group. If the values do not sum to 1, SYSTAT normalizes them so that they do. COLD.

© 1989, SYSTAT, Inc.

PROFILE=	Sets up contrasts between dependent variables for profile analysis. You must specify on of the following options. COLD.
DIFFERENCE I POLYNOMIAL	DIFFERENCE compares each within subjects measure with the following measure. POLYNOMIAL generates orthogonal polynomial comparisons. The default is equally spaced comparisons.
/METRIC=*m, n, ...*	Generates unequally spaced orthogonal polynomials. You must specify on number per dependent variable.
SUM	Computes and tests the sum of the dependent variables.
ROTATE=*n*	Rotates *n* factors by varimax rotation. COLD.
SAVE "*filename*"	Saves canonical scores and discriminant results. All standard SAVE options apply. COLD.
SINGLE I DOUBLE	Specifies whether to save in single or double precision.
STANDARDIZE=TOTAL I WITHIN	Specifies the method of standardizing canonical coefficients. TOTAL standardizes by the pooled variance from the entire sample. WITHIN, the default, standardizes by pooling variance within groups. COLD.
TYPE=CORRELATION I COVARIANCE I SSCP	Specifies the type of matrix to be factored. All factoring is done with listwise deletion. COLD.
TEST	Initiates the test of your hypothesis. HOT.

© 1989, SYSTAT, Inc.

Nonlin	**PRINT**=LONG I SHORT	LONG prints asymptotic standard errors and correlations for parameter estimates. SHORT prints default output. SAVE must be entered before ESTIMATE. COLD.
	SAVE "*filename*"	Saves residuals, estimated values, and the variables from MODEL in *filename*. All standard SAVE options apply. SAVE must be entered before ESTIMATE. HOT.
	SINGLE I DOUBLE	Specifies whether to save in single or double precision.
Loss function...	**LOSS=***function*	Specifies a loss function to apply in model estimation. The default is ordinary least squares. You may use the system variable ESTIMATE (the fitted values from your MODEL) in your loss function if it does not appear as a variable in your SYSTAT file. COLD.
Model...	**MODEL** *var=function*	Specifies a model to be estimated. Any terms which are not variables are assumed to be parameters. COLD.

© 1989, SYSTAT, Inc.

ESTIMATE	Tells SYSTAT to estimate your MODEL. HOT.
/ITER=*n*	Specifies the maximum number of iterations for fitting your model. The default is 20.
QUASI | SIMPLEX	Specifies the method of model estimation, QUASI, the default, specifies the Quasi-Newton method, while SIMPLEX specifies the Simplex method.
SCALE	Rescales the mean square error to 1 at the end of the iterations.
START=*n1, n2, ...*	Specifies starting values for MODEL parameters. The default is .1 for each parameter. You must specify starting values for all or none of your parameters. Specify starting values in the order the parameters appear in your MODEL and LOSS statements.
TOL=*n*	Specifies how closely fitted values must match actual values before the model is considered fit. The default is .001.

© 1989, SYSTAT, Inc.

Npar Friedman...	**FRIEDMAN** *varlist*	Computes a Friedman two-way analysis of variance on all variables in *varlist*. If you omit *varlist* all numeric variables are used. HOT.
Kolmogorov-Smirnov	**KS** *varlist* [* *gvar*]	Computes a Kolmogorov-Smirnov test. To compute a one-sample Kolmogorov-Smirnov test, follow *varlist* with one of the distribution options given below. To compute a two-sample test, specify only *varlist*; all variables in *varlist* must have equal *n* after deletion of missing values. For a two-sample test with unequal *n*, specify one numeric variable followed by a grouping variable. If you omit *varlist*, a two-sample test is computed using all numeric variables. HOT.
	/CHISQ=*n* | LILLIEFORS | NORMAL | UNIFORM	Specifies a distribution for the one-sample test. CHISQ=*n* compares the variables to a chi-square distribution with *n* degrees of freedom. LILLIEFORS compares the variables to a normal distribution with sample mean and standard deviation. NORMAL compares the variables to a standard normal distribution. UNIFORM compares the variables to a uniform distribution with mean 0 and standard deviation 1.
Kruskal-Wallis...	**KRUSKAL** *varlist* * *gvar*	Computes a Kruskal-Wallis analysis of variance on *varlist*. If *gvar* comprises only two levels, the Mann-Whitney (Wilcoxon) statistic is displayed. HOT.
Sign...	**SIGN** *varlist*	Computes a sign test on all pairs of variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.

© 1989, SYSTAT, Inc.

Wald-Wolfowitz runs...	**RUNS** *varlist*	Computes a Wald-Wolfowitz runs test on the numeric variables in *varlist*. If you omit *varlist*, all numeric variables are used. COLD.
	/CUT=*n*	Specify a baseline for determining runs. The default value is 0.
Wilcoxon...	**WILCOXON** *varlist*	Computes a Wilcoxon signed-rank test on all pairs of variables in *varlist*. If you omit *varlist*, all numeric variables are used. HOT.

© 1989, SYSTAT, Inc.

Series	**SAVE** "*filename*"	Saves the series resulting from the commands FOURIER, ARIMA, or SMOOTH into the SYSTAT file *filename*. COLD.
	SINGLE | DOUBLE	Specifies whether to save in single or double precision.
ARIMA...	**ARIMA** *var*	Estimates the Box-Jenkins ARIMA model. The default *var* is the first numeric variable in the file. HOT.
	/BACKAST=*n*	Specifies the number of backcasts. The default is 0, and *n* must be a positive integer.
	CONSTANT	Includes a constant in the ARIMA model. The constant is the sample mean.
	FORECAST =*n (or) time1-time2*	Specifies the number of forecasts or the time periods to forecast. *Time1*, *time2*, and *n* all must be positive integers.
	ITER=*n*	Specifies the maximum number of iterations allowed to fit your model. The default is 20.
	P=*n*	Specifies the number of AR parameters. the default is 0.
	PS=*n*	Specifies the number of seasonal AR parameters.
	Q=*n*	Specifies the number of MA parameters. the default is 0.
	QS=*n*	Specifies the number of seasonal MA parameters. the default is 0.
	SEASON=*n*	Specifies the seasonal periodicity. The default is 0.
	TOL=*n*	Specifies how closely the fitted values must match the actual values before estimates are considered converged. The default is 0.01. *N* must be greater than 0.

© 1989, SYSTAT, Inc.

Exponential...	**EXPONENTIAL** *var*	Specifies an exponential smoothing model. HOT.
	/ADDITIVE=*n*	Specifies an additive seasonal term. 0≤*n*≤1.
	FORECAST=*n* |*a*–*b*	Specifies the number of forecasts or a range of forecasts. *N* specifies the number of forecasts to follow the last case in the file. The range of cases is specified by *a* and *b*.
	LINEAR=*n*	Specifies linear growth. 0≤*n*≤1.
	MULTIPLICATIVE=*n*	Specifies multiplicative seasonal term. 0≤*n*≤1.
	PERCENTAGE=*n*	Specifies percentage growth. 0≤*n*≤1.
	SEASON=*n*	Specifies seasonal periodicity. *N* must be a positive integer.
	SMOOTH=*n*	Specifies weight for level term. 0≤*n*≤1.
Fourier...	**FOURIER**=*varlist*	Computes Fourier decomposition. HOT.

© 1989, SYSTAT, Inc.

LOWESS...	**SMOOTH** *var*	Smooths a series. HOT.
	/ LOWESS=*n*	Calculates Cleveland's smoother with parameter *n*, 0≤*n*≤1. The default is .5. As *n* increases, the function becomes smoother (tighter).
(Smooth...)	MEAN=*n*	Smooths the data with running means. *N* is the number of neighbors to include in the smoothed mean at a given point along the abscissa.
	MEDIAN=*n*	Smooths the data with running medians. *N* is the number of neighbors to include in the smoothed median at a given point along the abscissa.
	WT=*n*	Specifies the weighting to assign the neighbors of the data being smoothed. Specify an *n* for each neighbor included. The default is 1.0 for each neighbor.
Seasonal...	**ADJSEASON** *var*	Computes seasonal decomposition and adjustment on *var* by ratio-to-moving-averages method. The first numeric variable is used by default. HOT.
	/ADDITIVE | MULTIPLICATIVE	Specifies an additive or seasonal term.
	SEASON=*n*	Specifies seasonal periodicity *n*. The default value is 12.

© 1989, SYSTAT, Inc.

Case plot...	**PLOT**=*var*	Plot data values of *var*. HOT.
	/LAG=*n*	Specifies the number of cases to plot. The default is 15.
	MAX=*n*	Specifies the upper bound of the plot. The default is the data maximum.
	MIN=*n*	Specifies the lower bound of the plot. The default is the data minimum.
	STANDARDIZE	Standardizes *var* before plotting.
ACF plot...	**ACF** *var*	Graphs the autocorrelation function of *var*. The first numeric variable is the default. HOT.
	/LAG=*n*	Lags the data *n* cases before calculating the autocorrelation. The default is 15.
PACF plot...	**PACF** *var*	Graphs the partial autocorrelation function of *var*. The first numeric variable is the default. HOT.
	/LAG=*n*	Lags the data *n* cases before calculation the autocorrelation. The default is 15.
CCF plot...	**CCF** *var1, var2*	Calculates the cross correlation of the two numeric variables. You must specify both variables. HOT.
	/LAG=*n*	Lags the variable before calculating the cross correlation. This option produces *n*/2–1 positive and negative lags as well as LAG=0. If *n* is an even number, then one more negative lag is calculated. The default *n* is 15. *N* must be a positive integer.
Clear	**CLEAR** *var*	Clears *var* from the active workspace. All transformations are "undone." HOT.

© 1989, SYSTAT, Inc.

ID...	**ID** *n1, n2* [,*n3*]	Labels points in a series plot. *N1* is the year of the first observation in the series. *N2* is the periodicity. *N3* indicates the period of the first observation. COLD.
Transform...	**DIFFERENCE**=*var*	Transforms the data into differences between successive observations. HOT.
	/LAG=*n*	Specifies the seasonal period for the difference calculations. The default is 1. *N* must be a positive integer.
	INDEX=*var*	Replaces each value of *var* with the ratio of that value to the base observation. The default base is the first observation. HOT.
	/BASE=*n*	Specifies a base value. The default base is the first observation.
	LOG=*var*	Transforms the values of *var* into their natural logs. HOT.
	MEAN=*var*	Subtracts the mean of *var* from each observation. HOT.
	PCNTCHNG=*var*	Replaces each observation of *var* with the percentage change from the previous observation. HOT.
	SQUARE=*var*	Squares each observation of *var*. HOT.
	TAPER *var*	Split-cosine-bell tapers *var*. HOT.
	/P=*n*	Specifies the proportion of the series to be within the tapering window.
	TREND=*var*	Removes the linear trend from var. HOT.

© 1989, SYSTAT, Inc.

Stats	**PRINT**=LONG | <u>SHORT</u>	When calculating subgroup statistics (see the BY command) for two groups, LONG produces Bartlett's test and a *t*-test. For three or more groups, produces Bartlett's test and an ANOVA. Enter PRINT=LONG before the STATS command. SHORT produces default output. COLD.
	SAVE "*filename*"	Saves the first statistic listed as an option in the STATISTIC command into *filename*. All standard SAVE options apply. COLD.
	SINGLE | <u>DOUBLE</u>	Specifies whether to save in single or double precision.
Statistics...	**STATISTICS** *varlist*	Calculates specified statistics for each variable in *varlist*. HOT.
	/<u>MEAN MINIMUM</u> <u>MAXIMUM SD</u> CV KURTOSIS RANGE SEM SKEWNESS SUM VARIANCE	Mean, minimum, maximum, and standard deviation are calculated by default. If you specify statistics, only those that you specify are calculated.
	ALL	Calculates all eleven statistics.
(One-way ANOVA...)	TUKEY	Calculates Tukey HSD test. Precede STATISTICS with BY *gvar*.
T-test...	**TTEST** *varlist* [**gvar*]	Performs a dependent (paired samples) *t*-test on all pairs of variables in *varlist*. For an independent *t*-test, include the optional **gvar*, where *gvar* has only two values. HOT.

© 1989, SYSTAT, Inc.

Tables	**PRINT**=LONG | <u>SHORT</u>	LONG produces chi-square and PRE statistics when entered prior to TABULATE. SHORT produces default output. COLD.
	SAVE "*filename*"	Saves a table to *filename*. To use SAVE, you must subsequently TABULATE and MODEL. All standard SAVE options apply. COLD.
	/DIFF | FIT | RESID | <u>TABLE</u>	Specifies the statistic to save. DIFF saves residuals, FIT saves fitted values, RESID saves standardized residuals, and TABLE, the default, saves the table you TABULATE.
	SINGLE | <u>DOUBLE</u>	Specifies whether to save in single or double precision.
Tabulate...	**TABULATE** *varlist* * *var1* * *var2* * *varn*	Produces frequency tables for each variable in *varlist*. If you omit *varlist*, all variables are tabulated. To produce multiway tables, separate variables with asterisks (*). HOT.
	COLPCT	Produces column percent tables.
	CONFI=*n*	Prints confidence intervals on cell proportions at level n ($0 \leq n \leq 1$).
	<u>FREQUENCY</u>	Produces frequency tables.
	LIST	Produces list format tables.
	MISS	Excludes missing data from the table.
	PERCENT	Produces cell percent tables.
	COLPCT	Produces column percent tables.
	ROWPCT	Produces row percent tables.

© 1989, SYSTAT, Inc.

Log-linear models...	MODEL *constant* **varlist*, [* *var2* * *var3* * *varn*]	Fits a log-linear model to a table created with a prior TABULATE command. Constant is an optional term in the specification but the constant parameter is always part of the log-linear model. All variables specified in MODEL must have been tabulated. The model automatically includes a constant. All models are hierarchical; that is, all lower order effects are included in a model if higher order (interaction) effects are specified. For example, the term A*B*C encompasses the A*B*C effect, the two-way terms A*B, A*C, and B*C, and the terms A, B, and C. HOT.
	/CRITERION=*n*	Specifies how closely the fitted values must match the actual values before the model is considered fit. The default is .1.
	DELTA=*n*	Adds *n* to each cell count. This allows the inclusion of cells with sampling zeros. The default value is zero.
	DIFFERENCES	Produces a table of raw residuals.
	FITTED	Produces a table of fitted values. The default.
	ITERATIONS=*n*	Specifies the maximum number of iterations to fit the model. The default is 20.
	RESIDUALS	Produces a table of standardized residuals.
	ZERO	Defines all cells with zero observations as structural zeros.

© 1989, SYSTAT, Inc.

Appendix II: Command equivalents for examples

Mouse instructions are given for all examples in the manual. Here we provide the equivalent keyboard commands for those examples. Examples are numbered within each chapter; "4.3" means the third example of Chapter 4.

2 Cluster

2.1 Joining with Euclidean distances

```
USE DATAFILE
JOIN ACCIDENT-LIVER/COLS
```

2.2 Joining with correlation metric

```
USE MEDICAL
DISTANCE=PEARSON
JOIN ACCIDENT,CANCER,CARDIO,LIVER,
  PNEU_FLU,PULMONAR,DIABETES/COLS
```

2.3 Joining a matrix

```
USE WINES
LABEL YEAR$
JOIN/MATRIX,SYMBOL=7
```

2.4 K-means

```
USE DATAFILE
LABEL STATE$
KMEANS ACCIDENT,CANCER,CARDIO,LIVER,
  PNEU_FLU,PULMONAR,DIABETES/NUMBER=3
```

2.5 Distance matrix input

```
USE DATAFILE
PRINT=LONG
JOIN BERLIN BOMBAY CAPETOWN CHICAGO LONDON,
   MONTREAL NEW YORK PARIS SANFRAN SEATTLE
```

© 1989, SYSTAT, Inc.

3 Corr

3.1 Pearson correlation

```
USE DATAFILE
PEARSON ACCIDENT-LIVER
```

3.2 Probabilities

```
USE DATAFILE
PEARSON ACCIDENT-LIVER/PROB
```

3.3 Bonferroni-adjusted probabilities

```
USE DATAFILE
PEARSON ACCIDENT-LIVER/BPROB
```

4 Factor

4.1 Principal components

```
USE DATAFILE
SAVE MEDSCORE
NUMBER=2
ROTATE=VARIMAX
PRINT=LONG
SORT
FACTOR ACCIDENT,CARDIO,CANCER,PULMONAR,
  PNEU_FLU,DIABETES,LIVER/PLOT
```

4.2 Plot factor scores

```
DATA
USE MEDSCOR MEDICAL
SAVE MED
LET RURALURB=FACTOR(1)
LET CLIMATE=FACTOR(2)
RUN

GRAPH
PLOT CLIMATE*RURALURB/SYMBOL=STATE$
```


© 1989, SYSTAT, Inc.

Appendix II: Command equivalents for examples

5 MDS

5.1 Guttman method

```
USE EKMAN
METHOD=GUTTMAN
SCALE
```

5.2 Kruskal method

```
USE ROTHKOPF
SHEPARD ?
SCALE
```

6 MGLH

6.1 Simple regression

```
USE ROTATE
MODEL RT=CONSTANT+ANGLE
ESTIMATE
```

6.2 Multiple regression

```
USE LONGLEY
MODEL TOTAL=CONSTANT+DEFLATOR+GNP+UNEMPLOY,
  +ARMFORCE+POPULATN+TIME
ESTIMATE

MODEL TOTAL=CONSTANT+DEFLATOR+GNP+UNEMPLOY,
  +ARMFORCE+POPULATN+TIME
PRINT=LONG
ESTIMATE
```

6.3 Testing hypothesis

```
HYPOTHESIS
EFFECT=DEFLATOR
TEST

HYPOTHESIS
EFFECT=DEFLATOR & GNP
TEST
```

© 1989, SYSTAT, Inc.

6.4 Regression with ecological or grouped data

```
USE PLANTS
WEIGHT=COUNT
MODEL CO2=CONSTANT+SPECIE
ESTIMATE
```

6.5 Weighted least squares

```
USE NIELSEN
MODEL Y=CONSTANT+X
SAVE RESID
ESTIMATE

USE NIELSEN
MODEL WY=WEIGHT+WX
SAVE RESID
ESTIMATE
```

6.7 Mixture models

```
USE PUNCH
MODEL VALUE=W+P+O+W*P+W*O+P*O
ESTIMATE/MIX
```

6.8 Polynomial regression

```
USE USPOP
MODEL POP=CONSTANT+YEAR+YEAR*YEAR
ESTIMATE
```

6.9 Stepwise regression

```
USE LONGLEY
MODEL TOTAL=CONSTANT+DEFLATOR+GNP+UNEMPLOY,
  +ARMFORCE+POPULATN+TIME
STEP
```

6.10 Partial correlations

```
PRINT=LONG
MODEL Y(1-5)=X
ESTIMATE
```


© 1989, SYSTAT, Inc.

Appendix II: Command equivalents for examples

6.11 Analysis of variance

```
USE KENTON
ANOVA SALES
CATEGORY PACKAGE=4
ESTIMATE


USE TEMP
FORMAT=0
LIST SALES,X(1-3)
RUN

HYPOTHESIS
EFFECT=PACKAGE
CONTRAST
1 0 -1 0
TEST

HYPOTHESIS
EFFECT=PACKAGE
CONTRAST
1 1 1 -3
TEST
```

6.12 Two-way analysis of variance

```
USE AFIFI
CATEGORY DRUG=4,DISEASE=3
ANOVA SYSINCR
SAVE RESID
ESTIMATE

HYPOTHESIS
EFFECT=DRUG
CONTRAST
1 0 -1 0
TEST

HYPOTHESIS
EFFECT=DRUG
CONTRAST
1 0 0 -1
TEST
```

© 1989, SYSTAT, Inc.

6.13 Single degree of freedom designs

```
USE REACT
CATEGORY A=2,B=2,C=2,D=2
ANOVA YIELD
ESTIMATE

SAVE EFFECTS/COEF
ESTIMATE

DATA
USE EFFECTS (X(1-15))
SAVE EFFECTS2
TRANSPOSE
RUN

GRAPH
USE EFFECTS2
PPLOT
```

6.15 Randomized block designs

```
USE BLOCK
CATEGORY BLOCK=5 TREAT=3
MODEL JUDGEMENT=CONSTANT+BLOCK+TREAT
ESTIMATE
```

6.16 Incomplete block designs

```
USE JOHN
CATEGORY BLOCK=8,A=2,B=2,C=2
MODEL Y=CONSTANT+BLOCK+A+B+C+A*B+A*C+B*C+A*B*C
ESTIMATE
```

6.17 Fractional factorial designs

```
USE FRACTION
CATEGORY A=2,B=2,C=2,D=2
MODEL Y=CONSTANT+A+B+C+D+A*B+A*C+B*C
ESTIMATE
```

6.18 Mixed models

```
USE AFIFI
CATEGORY DRUG=4,DISEASE=3
ANOVA SYSINCR
ESTIMATE
```


© 1989, SYSTAT, Inc.


```
HYPOTHESIS
EFFECT=DRUG
ERROR=DRUG*DISEASE
TEST
```

6.19 Nested designs

```
MGLH
USE SCHOOLS
CATEGORY TEACHER(1)=2,TEACHER(2)=2,
  TEACHER(3)=2,SCHOOL=3
MODEL LEARNING=CONSTANT+TEACHER(1)+TEACHER(2),
  +TEACHER(3)+SCHOOL
ESTIMATE

HYPOTHESIS
EFFECT=TEACHER(1) & TEACHER(2) & TEACHER(3)
TEST
```

6.20 Split plot designs

```
USE PLOTS
CATEGORY PLOT(1)=2,PLOT(2)=2,A=2,B=4
MODEL YIELD=CONSTANT+A+B+A*B+PLOT(1)+PLOT(2),
  +B*PLOT(1)+B*PLOT(2)
ESTIMATE

HYPOTHESIS
EFFECT=A
ERROR=PLOT(1) & PLOT(2)
TEST

HYPOTHESIS
EFFECT=B
ERROR=B*PLOT(1) & B*PLOT(2)
TEST

HYPOTHESIS
EFFECT=A*B
ERROR=B*PLOT(1) & B*PLOT(2)
TEST
```

6.21 Latin square designs

```
USE LATIN
CATEGORY DAY=5,WEEK=5,SQUARE=5
MODEL RESPONSE=CONSTANT+DAY+WEEK+SQUARE
ESTIMATE
```

© 1989, SYSTAT, Inc.

6.22 Crossover and changeover designs

```
CATEGORY COW=6,FEED=3,RESIDUAL=3,PERIOD1=3,PERIOD2=3
MODEL MILK=CONSTANT+COW+PERIOD1+PERIOD2+,
  FEED+RESIDUAL
```

6.23 Analysis of covariance

```
USE COVAR
CATEGORY TREAT=3
MODEL Y=CONSTANT+TREAT+X+TREAT*X
ESTIMATE

MODEL Y=CONSTANT+TREAT+X
ESTIMATE
```

6.24 Hotellings T^2

```
MODEL X,Y=CONSTANT
ESTIMATE
HYPOTHESIS
TEST

HYPOTHESIS
DMATRIX
10 15
TEST

CATEGORY GROUP=2
MODEL X,Y=CONSTANT+GROUP
ESTIMATE

HYPOTHESIS
EFFECT=GROUP
TEST
```


© 1989, SYSTAT, Inc.

6.25 Multivariate analysis of variance

To specify a fully factorial ANOVA model, you can use either the MODEL or the ANOVA command.

```
USE MANOVA
CATEGORY,
  SEX=2,DRUG=3
MODEL WEEK(1-2)=CONSTANT,
  +SEX+DRUG+SEX*DRUG
ESTIMATE

HYPOTHESIS
EFFECT=SEX
TEST
PRINT=LONG
HYPOTHESIS
EFFECT=DRUG
TEST
PRINT=SHORT
HYPOTHESIS
EFFECT=SEX*DRUG
TEST
```

or

```
USE MANOVA
CATEGORY,
  SEX=2,DRUG=3
ANOVA WEEK(1-2)
ESTIMATE
```

6.27 Multivariate profile analysis

```
USE RATS
MODEL WEIGHT(1-5)=CONSTANT
ESTIMATE

HYPOTHESIS
PROFILE=POLYNOMIAL
TEST
```

Unevenly spaced polynomials

```
HYPOTHESIS
PROFILE=POLYNOMIAL/METRIC=1,2,3,4,10
TEST
```

Difference contrasts

```
HYPOTHESIS
PROFILE=DIFFERENCE
TEST
```

© 1989, SYSTAT, Inc.

Summing effects

```
HYPOTHESIS
PROFILE=SUM
TEST
```

6.28 Unweighting means

```
USE MOTHERS
CATEGORY CLASS=4
MODEL SCALE(1-3)=CONSTANT+CLASS
ESTIMATE

HYPOTHESIS
PROFILE=DIFFERENCE
EFFECT=CONSTANT
TEST

HYPOTHESIS
PROFILE=DIFFERENCE
AMATRIX
21 4 1 0
TEST
```

6.29 One-way repeated measures

To specify a fully factorial repeated measures model, you can use either the MODEL or the ANOVA command.

```
USE RATS                          or    USE RATS
MODEL WEIGHT(1-5)=CONSTANT,             ANOVA WEIGHT(1-5),
 /REPEAT                                 /REPEAT
ESTIMATE                                ESTIMATE
```

6.30 One grouping factor

To specify a fully factorial repeated measures model, you can use either the MODEL or the ANOVA command.

```
USE REPEAT1                       or    USE REPEAT1
CATEGORY ANXIETY=2                      CATEGORY ANXIETY=2
MODEL TRIAL(1-4)=,                      ANOVA TRIAL(1-4),
  CONSTANT+ANXIETY/REPEAT                /REPEAT
ESTIMATE                                ESTIMATE
```


© 1989, SYSTAT, Inc.

Two grouping factors

To specify a fully factorial repeated measures model, you can use either the MODEL or the ANOVA command.

```
USE REPEAT1                       or    USE REPEAT1
CATEGORY,                               CATEGORY,
 ANXIETY=2,                              ANXIETY=2,
  TENSION=2                               TENSION=2
MODEL TRIAL(1-4)=,                      ANOVA TRIAL(1-4),
  CONSTANT+ANXIETY+TENSION,               /REPEAT
  +ANXIETY*TENSION/REPEAT  ESTIMATE
ESTIMATE
```

6.31 Repeated measures with two trial factors

To specify a fully factorial repeated measures model, you can use either the MODEL or the ANOVA command.

```
USE REPEAT2                       or    USE REPEAT2
CATEGORY NOISE=2                        CATEGORY NOISE=2
MODEL P1D1-P3D3=,                       ANOVA P1D1-P3D3,
  CONSTANT+NOISE,                        /REPEAT=3,3
   /REPEAT=3,3                          ESTIMATE
ESTIMATE
```

6.32 Repeated measures with three or more trial factors

```
CATEGORY D=2
REP1, REP2, REP3, REP4, REP5, REP6, REP7, REP8=D
ESTIMATE

HYPOTHESIS
CMATRIX
-1 -1 -1 -1 1 1 1 1
ALL
TEST

HYPOTHESIS                        or    HYPOTHESIS
CMATRIX                                 PROFILE=SUM
1 1 1 1 1 1 1 1                         TEST
EFFECT=D
TEST
```

© 1989, SYSTAT, Inc.

```
HYPOTHESIS
PROFILE=SUM
EFFECT=0
TEST
```

6.33 Repeated measures analysis of covariance

To specify a fully factorial repeated measures analysis of covariance model, you can use either the MODEL or the ANOVA and COVARIATE commands.

```
USE WINER                          or      USE WINER
CATEGORY SEX=2                             CATEGORY SEX=2
MODEL DAY(1-2)=,                           ANOVA DAY(1-
2)/REPEAT
  CONSTANT+SEX+AGE/REPEAT                  COVARIATE AGE
ESTIMATE                                   ESTIMATE
```

6.34 Covariance alternatives to repeated measures

```
CATEGORY GROUP=2
MODEL POST=CONSTANT+GROUP+PRE
ESTIMATE

CATEGORY GROUP=2
MODEL POST=CONSTANT+GROUP+PRE+GROUP*PRE
ESTIMATE
```

6.35 Issues in repeated measure analysis

```
USE REPEAT1
CATEGORY ANXIETY=2, TENSION=2
MODEL TRIAL(1-4)=CONSTANT+ANXIETY+TENSION+,
  ANXIETY*TENSION/REPEAT
ESTIMATE
HYPOTHESIS
EFFECT=ANXIETY
PROFILE=DIFFERENCE
TEST
```


© 1989, SYSTAT, Inc.

6.36 Discriminant analysis

```
MGLH
USE IRIS
CATEGORY SPECIES=3
MODEL SEPALLEN,SEPALWID,PETALLEN,PETALWID=,
  CONSTANT+SPECIES
PRINT LONG
ESTIMATE

HYPOTHESIS
EFFECT=SPECIES
SAVE CANON
TEST
DATA
USE CANON
SAVE DISCRIM
LABEL GROUP/1='A',2='B',3='C'
LIST
RUN

TABLES
USE DISCRIM
TABULATE GROUP*PREDICT

GRAPH
USE DISCRIM
PLOT FACTOR(2)*FACTOR(1)/SYMBOL=GROUP$
```

6.37 Principal components analysis

```
MGLH
USE MEDICAL
CATEGORY DIVISION=9
MODEL ACCIDENT,CARDIO,CANCER,PULMONAR,
  PNEU_FLU,DIABETES,LIVER=CONSTANT+DIVISION
ESTIMATE

HYPOTHESIS
EFFECT=DIVISION
FACTOR=ERROR
TYPE=CORRELATION
ROTATE=2
TEST
```

© 1989, SYSTAT, Inc.

6.38 Canonical correlation analysis

```
MODEL MMPI(1-10) =CONSTANT +RATER(1)+ RATER(2),
  +RATER(3)
ESTIMATE
PRINT=LONG
HYPOTHESIS
STANDARDIZE=TOTAL
EFFECT=RATER(1) & RATER(2) & RATER(3)
TEST

MODEL RATER(1-3)=CONSTANT+MMPI(1)+MMPI(2),
  +MMPI(3)+MMPI(4)+MMPI(5)+MMPI(6)+MMPI(7),
  +MMPI(8)+MMPI(9)+MMPI(10)
ESTIMATE

HYPOTHESIS
STANDARDIZE=TOTAL
EFFECT=MMPI(1) & MMPI(2) & MMPI(3) & MMPI(4) &,
  MMPI(5) & MMPI(6) & MMPI(7) & MMPI(8) &,
  MMPI(9) & MMPI(10)
TEST
```

7 Nonlin

7.1 Estimating a mean

```
USE IRIS
PRINT=LONG
SELECT SPECIES=1
MODEL SEPALLEN=MEAN
ESTIMATE
```

7.2 Estimating a median

```
USE IRIS
SELECT SPECIES=1
MODEL SEPALLEN=MEDIAN
LOSS=ABS(SEPALLEN-ESTIMATE)
ESTIMATE/SIMPLEX
```

7.3 Linear regression

```
USE ROTATE
PRINT=LONG
MODEL RT=CONSTANT+BETA*ANGLE
ESTIMATE
```


© 1989, SYSTAT, Inc.

7.4 Least absolute values regression

```
USE ROTATE
MODEL RT=CONSTANT+BETA*ANGLE
LOSS=ABS(RT-ESTIMATE)
ESTIMATE/SIMPLEX
```

7.5 Nonlinear regression

```
USE USPOP
PRINT=LONG
MODEL POP=EXP(C+B1*YEAR)
ESTIMATE/START=0,0

SAVE RESID

GRAPH
USE RESID
PLOT ESTIMATE*YEAR

PLOT RESIDUAL*ESTIMATE

PPLOT RESIDUAL

SERIES
USE RESID
ACF RESIDUAL
```

7.6 Piecewise regression

```
USE LEARN
PRINT=LONG
MODEL TASKS=(B*SESSION^2)*(SESSION<KNOW),
  +(B*KNOW^2)*(SESSION=>KNOW)
ESTIMATE/START=.02,8
```

7.7 Maximum likelihood estimation

```
USE IRIS
PRINT=LONG
SELECT SPECIES=1
MODEL SEPALLEN=MEAN
LOSS=-.5*LOG(2*3.14159)+.5*LOG(VARIANCE)+,
  (1/(2*VARIANCE))*(SEPALLEN-ESTIMATE)^2
ESTIMATE/SIMPLEX,SCALE
```

© 1989, SYSTAT, Inc.

7.8 Logistic models—iteratively reweighted least squares

```
USE COX
MODEL FAILURE=COUNT*EXP(B0+B1*TIME),
 /(1+EXP(B0+B1*TIME))
LOSS=(FAILURE-ESTIMATE)^2*COUNT,
 /(ESTIMATE*(COUNT-ESTIMATE))
ESTIMATE/SCALE
```

7.9 Logistic models—maximum likelihood estimation

```
USE COX2
MODEL FAILURE=EXP(B0+B1*TIME)/(1+EXP(B0+B1*TIME))
LOSS=-COUNT*(FAILURE*LOG(ESTIMATE),
  +(1-FAILURE)*LOG(1-ESTIMATE))
FORMAT 6
ESTIMATE/SCALE
```

7.10 Probit models—maximum likelihood estimation

```
USE HABERMAN
PRINT=LONG
MODEL RESPONSE=ZCF(B0+B1*YEAR)
LOSS=-COUNT*(RESPONSE*LOG(ESTIMATE),
  +(1-RESPONSE)*LOG(1-ESTIMATE))
ESTIMATE/SCALE

SAVE RESID

GRAPH
USE RESID
PLOT ESTIMATE*YEAR
```

7.11 Kinetic models

```
USE ENZYME
PRINT=LONG
MODEL V=VMAX*S/(KM*(1+I/KIS)+S)
ESTIMATE
```

7.12 Deterministic models—minimizing an analytic function

```
DATA
SAVE DUMMY
REPEAT 1
LET DUMMY=0
RUN
```


© 1989, SYSTAT, Inc.


```
NONLIN
USE DUMMY
MODEL DUMMY=DUMMY
LOSS=100*(U-V^2)^2+(1-V)^2
ESTIMATE/SIMPLEX
```

7.13 A random model

```
USE MEDICAL
MODEL CARDIO=BETA*URN
ESTIMATE
```

7.14 Contouring the loss function

```
PLOT
LOSS*THETA2*THETA1/XTICK=10,YTICK=10,ZTICK=1,ZMIN=0,
ZMAX=10000
```

8 Npar

8.1 Sign test

```
USE MOTHERS
SIGN MOTHERS(1-2)
```

8.1 Friedman two-way analysis of variance

```
USE MOTHERS
FRIEDMAN
```

8.2 Kolmogorov-Smirnov one-sample test

```
DATA
REPEAT 50
SAVE NORM
LET Z=ZRN
LET ZS=Z
STANDARDIZE ZS
RUN

NPAR
USE NORM
KS/NORMAL
```

© 1989, SYSTAT, Inc.

8.3 Lilliefors test

```
USE NORM
KS Z,ZS/LILLIEFORS
```

8.4 Kolmogorov-Smirnov two-sample test

```
USE NORM
KS Z,ZS
```

8.5 Kruskal-Wallis one-way analysis of variance

```
USE CRIMINAL
CATEGORY DIVISION=9
MODEL PRISONER=CONSTANT+DIVISION
ESTIMATE

USE CRIMINAL
KRUSKAL PRISONER*DIVISION
```

8.7 Wald-Wolfowitz runs test

```
USE SALES
RUNS MRCHDSE/CUT=1743

USE SALES
RUNS MRCHDSE/CUT=2517
```

8.8 The Wilcoxon signed ranks test

```
USE MOTHERS
WILCOXON MOTHERS(1-2)
```

9 Series

9.1 Series plots and differencing

```
USE SPNDMONY
ID 1952,4
PLOT SPENDING/LAG=20

DIFFERENCE SPENDING
PLOT SPENDING/LAG=20
PLOT MONEY
```


© 1989, SYSTAT, Inc.

9.2 ACF, PACF, CCF plots

```
USE SPNDMNY
ACF SPENDING
DIFFERENCE SPENDING
ACF SPENDING

CLEAR
PACF SPENDING

CCF SPENDING,MONEY
DIFFERENCE SPENDING
DIFFERENCE MONEY
CCF SPENDING, MONEY
```

9.3 ARIMA Model identification

```
USE AIRLINE
ID 1949,12
PLOT/LAG=144

LOG PASS

ACF
PACF

DIFFERENCE
ACF
PACF

DIFFERENCE/LAG=12
ACF
PACF
```

9.4 ARIMA with backcasting

```
ARIMA/Q=1, QS=2, SEASON=12, BACKCAST=13
```

9.5 ARIMA with forecasting

```
ARIMA/Q=1,QS=1,SEASON=12,BACKCAST=13,FORECAST=10
```

© 1989, SYSTAT, Inc.

9.6 Exponential smoothing

```
USE AIRLINE
EXPONENTIAL/SMOOTH=.3,LINEAR=.4,MULT=.4,FORECAST=10

EXPONENTIAL/SMOOTH=.2,LINEAR=.2,MULT=.2

EXPONENTIAL/SMOOTH=.3,LINEAR=.4,MULT=.4,
  FORECAST=121-144
```

9.7 Seasonal decomposition and adjustment

```
USE AIRLINE
ID 1949,12
PLOT/LAG=144

ADJSEASON PASS/MULT
ADJSEASON PASS/ADDITIVE
PLOT/LAG=144
```

10 Stats

10.1 Basic statistics

```
USE MEDICAL
STATISTICS ACCIDENT-LIVER
```

10.2 Statistics by groups

```
USE IRIS
BY SPECIES
PRINT LONG
STATISTICS/TUKEY
```

10.3 Two sample *t*-test

```
USE TRAND
TTEST X*GROUP
```

10.4 Dependent *t*-test

```
USE TRAND
TTEST X1,X2
```


© 1989, SYSTAT, Inc.

Appendix II: Command equivalents for examples

10.5 Analysis of variance

```
USE IRIS
BY SPECIES
PRINT=LONG
STATS SEPALLEN,SEPALWID,PETALLEN,PETALWID/TUKEY
```

11 Tables

11.1 One-way frequency table

```
USE DATAFILE
TABULATE REGION
```

11.2 List table

```
TABULATE REGION$/LIST
```

11.3 Confidence intervals

```
USE BUSES
TABULATE DISTANCE/PERCENT,CONFI=.95
```

11.4 Two-way table with tests and association measures

```
USE DANISH
WEIGHT=COUNT
FORMAT=0
TABULATE FATHER*SON
```

11.5 Percents

```
TABULATE FATHER*SON/PERCENT,ROWPCT,COLPCT
```

© 1989, SYSTAT, Inc.

11.6 Mantel-Haenszel test

```
USE ANSFIELD
WEIGHT=NUMBER
PRINT=LONG
TABULATE CANCER$*TREAT$*RESPONSE$/SORT
```

11.7 Log-linear model

```
USE DANISH
WEIGHT=COUNT
TABULATE FATHER*SON
MODEL FATHER+SON/FITTED,RESIDUALS
```

11.8 Log-linear models with multi-way tables

```
USE CANCER
WEIGHT=NUMBER
TABULATE CENTER$*AGE*SURVIVE$*INFLAPP$/SORT
MODEL AGE*CENTER$+INFLAPP$*CENTER$+SURVIV$*CENTER,
 /FITTED,DELTA=.5
```


© 1989, SYSTAT, Inc.

Appendix III: Data files

All the datasets used in the "Examples" sections of this manual are listed below. In the left margin are the filename and reference for each dataset. The text gives background information and details about variables.

AFIFI
Afifi and Azen (1974)

The dependent variable, SYSINCR, is the increase in systolic blood pressure after administering one of four different drugs (DRUG) to patients with one of three different diseases (DISEASE). Patients were assigned randomly to one of the four possible drugs.

AIRCRAFT
Bennet, Desmarais (1975)

These data show amplitude of vibration (FLUTTER) versus time (T) in an aircraft wing component.

AIRLINE
Box, Jenkins (1976)

The variable PASS contains monthly totals of international airline passengers for twelve years beginning in January, 1949.

ANSFIELD
Ansfield, et. al., (1977)

This study examined the effects (RESPONSE$) of treatments (TREAT$) on two patient groups (CANCER$): those with cancer of the colon or rectum and those with breast cancer. NUMBER gives the number of patients in each cancer/treatment/response group.

BLOCK
Neter, Wasserman, Kutner (1985)

These data comprise a randomized block design. Five blocks of judges (BLOCK) analyzed three treatments (TREAT). Subjects (judges) are stratified within blocks, so the interaction of blocks and treatments cannot be analyzed.

BUSES
Davis (1952)

These data count the number of buses failing (COUNT) after driving one of 10 distances (DISTANCE)

CANCER
Morrison (1973); Bishop, Fienberg, Holland (1975)

This study examined breast cancer patients in 3 diagnostic centers (CENTER$), 3 age groups (AGE), whether they survived after three years post-diagnosis (SURVIVE$), and the inflammation type (minimum/ maximum) and appearance of the tumor (malignant/ benign), combined in the variable INFLAPP$.

© 1989, SYSTAT, Inc.

COVAR
Winer (1971)

Winer uses this artificial dataset in an analysis of covariance in which Y is the dependent variable, X is the covariate, and TREAT is the treatment.

COX
Cox (1970)

These data record tests for failures among objects after certain times (TIME). FAILURE is the number of failures and COUNT is the total number of tests.

COX2
Cox (1970)

These data are a different form of the Cox dataset. Here, FAILURE is binary ("1" indicates failures, and "2" indicates successes), and COUNT is the number of failures or successes.

CRIME
US Statistical Abstract (1970); Bureau of the Census

These are crime rates per 100,000 people in each state.

DANISH
Svalastoga (1959)

The variables FATHER and SON represent social class indices of fathers and sons, respectively. COUNT gives the number of cases in each FATHER/SON combination.

EKMAN
Ekman (1954)

These data are judged similarities among 14 different spectral colors (The variable names are the colors' wavelengths). The judgements are averaged across 31 subjects.

ENZYME
Greco, et al. (1982)

These data measure competitive inhibition for an enzyme inhibitor. V is the initial enzyme velocity, X is the concentration of the substrate, and I is the concentration of the inhibitor.

FRACTION

These data comprise a fractional factorial design where data appear in only 8 out of 16 possible cells. Each cell contains 2 cases. Four treatment factors (A, B, C, and D) predict one dependent variable (Y).

HABERMAN
Haberman (1982)

These data represent the number of men at different levels of education (YEAR) who believe that women should stay at home rather than work. "0" represents an opinion that women should stay at home and "1" represents an opposite opinion.

© 1989, SYSTAT, Inc.

IRIS
Anderson (1935)

These data measure sepal length (SEPALLEN), sepal width (SEPALWID), petal length (PETALLEN), and petal width (PETALWID) in centimeters for three SPECIES of irises (1=Setosa, 2=Versicolor, and 3=Virginica).

JOHN
John (1971)

These data comprise an incomplete block design with three treatment factors (A, B, and C), a blocking variable with eight levels (BLOCK), and the dependent variable (Y).

KENTON
Neter, Wasserman, Kutner (1985)

These data comprise unit sales of a product (SALES) under different types of package designs (PACKAGE). Each case represents a different store.

LATIN
Neter, Wasserman, Kutner (1985)

These data comprise a Latin square design in which the response (RESPONSE) of a different square (SQUARE) was tested five days a week (DAY) for five weeks (WEEK).

LEARN
Gilfoil (1982)

These data demonstrate a quadratic function with a ceiling. They are from a study showing that inexperienced computer users prefer dialogue menu interfaces while experienced users prefer command based interfaces. SESSION is the session number and TASKS is the number of command-based (as opposed to dialogue-based) tasks initiated by the user during that session.

LONGLEY
Longley (1967)

Economic data selected by Longley to illustrate computational shortcomings of statistical software.

MANOVA
Morrison (1976)

These data comprise a hypothetical experiment measuring weight loss in rats. Each rat was assigned randomly to one of 3 drugs (DRUG), with weight loss measured in grams for the first and second weeks of the experiment (WEEK(1-2)).

MEDICAL
State and Metropolitan Area Data Book, 1986; Bureau of the Census

These data are death rates by cause from 9 census divisions of the U.S. in 1986.

MOTHER
Dixon (1981)

In this dataset, 20 mothers (MOTHERS(1–20) were ranked by 13 judges (13 cases).

© 1989, SYSTAT, Inc.

MOTHERS
Morrison (1976)

These data are hypothetical profiles on 3 scales of mothers (SCALE(1-3)) in each of 4 socioeconomic classes (CLASS).

NEWARK
U.S. Government

These data are 64 average monthly temperatures in Newark, New Jersey, beginning with January 1964. The data were collected by the U.S. Government and cited in Chambers, et al. (1983).

NIELSON
Neter, Wasserman, Kutner (1985)

These are weighted least squares data comparing the size of a construction bid in millions of dollars (X) and the cost to a construction firm to prepare the bid in thousands of dollars (Y).

PUNCH
Cornell (1985)

These data measure the effects of various mixtures of watermelon (W), pineapple (P), and orange juice (O), on test ratings by judges (VALUE) of a fruit punch.

RATS
Morrison (1976)

For these data, 6 rats were weighed at the end of each of 5 weeks (WEIGHT(1-5).

REACT
SYSTAT

These data describe yields of a chemical reaction (YIELD) under various combinations of four binary factors (A,B,C, and D). Two reactions were observed under each combination of experimental factors, so each cell has two cases.

REPEAT1
Winer (1971)

These data contain 2 grouping factors (ANXIETY and TENSION) and one trials factor (TRIAL(1-4)).

REPEAT2
Winer (1971)

This dataset has one grouping factor (NOISE) and two trials factors: PERIOD and DIAL. The trials factors must be entered as dependent variables in a MODEL statement, so we name the variables P1D1, P1D2, ..., P3D3. For example, P1D2 means a score in the (period1, dial2) cell.

ROTATE
Metzler and Shepard (1974)

These data measure reaction time in seconds (RT) versus angle of rotation in degrees (ANGLE) in a perception study. The experiment measured the time it took subjects to make "same" judgements when comparing a picture of a three-dimensional object to a picture of possible rotations of the object.

© 1989, SYSTAT, Inc.

ROTHKOPF
Rothkopf (1957)

These data are adapted from an experiment by Rothkopf in which 598 subjects were asked to judge whether Morse code signals presented two in succession were the same. All possible ordered pairs were tested. For multidimensional scaling, the data for letter signals have been averaged across sequence and the diagonal (pairs of the same signal) has been omitted.

SCHOOLS
Neter, Wasserman, Kutner (1985)

These data comprise a nested design where 2 teachers from each of 3 different schools were rated. SCHOOL indicates the school that the case describes. Each teacher variable (TEACHER(1-3)) represents a different school; a value of "1" indicates teacher 1 for that school, "2" indicates teacher 2 for that school, and "0" indicates that the teacher does not teach at that school. LEARNING measures the teachers effectiveness (the higher, the better).

SPNDMONY
Chatterjee, Price (1977)

In this dataset, SPENDING is consumer expenditures and MONEY is money stock in billions of dollars in each quarter of the years 1952–1956.

TRAND
SYSTAT

The variable X was generated by SYSTAT's normal random number generator. X1 contains the values of X on cases 1–10, and X2 contains the values of X on cases 11–20.

USDATA
State and Metropolitan Area Data Book (1986)

REGION and REGION$ divide the country into four regions. DIVISION and DIVISION$ divide the country into nine regions.

LANDAREA is in square miles, 1980. POP85 is the 1985 population in thousands.

ACCIDENT is the number of deaths by accident per 100,000 people. CARDIO is the number of deaths from major cardio-vascular disease per 100,000 people. CANCER is the number from cancer per 100,000 people. PULMONAR is the number from chronic obstructive pulmonary disease per 100,000 people. PNEU_FLU is the number from pneumonia and influenza per 100,000 people. DIABETES is the number from diabetes melitus per 100,000 people. LIVER is the number from chronic liver disease and cirrhosis per 100,000 people.

DOCTORS is the number of active, non-Federal physicians practicing in 1982.

HOSPITAL is the number of hospitals in 1983.

© 1989, SYSTAT, Inc.

MARRIAGE is the number of marriages in thousands in 1983. DIVORCE is the number of divorces and annulments in thousands in 1983.

FDSTAMPS is the number of participants in the Federal food stamp program in thousands in 1983.

TEACHERS is the number of teachers in thousands and TCHRSAL is the average salary for teachers for the 1983-84 school year. HSGRAD is the number of public high school graduates in the 1982-83 school year.

AVGPAY is the average annual pay for a worker in 1983.

The following are sales in millions of dollars for establishments with payrolls in 1982: TOTALSLE is the total sales; BLDGMTRL is building materials, hardware, garden supplies, and mobile homes; MRCHDSE includes general merchandise stores; FOODSTRS includes food stores; AUTODLRS includes automotive dealers; GASSTATN includes gasoline service stations; APPAREL includes apparel and accessory stores; FURNITUR includes furniture, home furnishing, and equipment stores; EATNDRNK includes eating and drinking establishments; DRUGSTRE includes drug and proprietary stores.

VIOLENT is the number in thousands of violent crimes in 1983. VIOLRATE is the violent crime rate rate per 100000 people in 1983. PROPERTY is the number in thousands of property crimes in 1983. PROPRATE is the rate of property crimes per 100000 in 1983. PRISONER is the total number of federal and state prisoners in mid-1984.

SPIRITS, WINE, and BEER are the average number of gallons purchased per person annually. (Source: Distilled Spirits Council of U.S.; Wine Institute; U.S. Brewers Association; Bureau of the Census, Department of Commerce—USA Today.)

TAXES is the total taxes paid for each $1 returned in federal aid. (Source: Tax Foundation—Chicago Tribune.)

© 1989, SYSTAT, Inc.

WILLIAMS
Cochran and Cox (1957)

These data comprise a crossover design for an experiment studying the effect of 3 different feed schedules (FEED) on milk production by cows (MILK). The design of the study has the form of two 3 x 3 Latin squares. PERIOD1 represents period within the first square, and PERIOD2 represents period within the second square. RESIDUAL indicates the treatment of the preceding period.

WINES
Simon (1969)

These data were prepared by the International Wine and Food Society. They show the ratings of seven types of wines (0=poor, 7=outstanding) over a 20 year period.

WINER
Winer (1971)

This design has two trials (DAY(1–2), one covariate (AGE), and one grouping factor (SEX).

USPOP
1987 Information Please Almanac

These data are the U.S. population in thousands (POP) beginning in 1790.

© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

Appendix IV: References

Afifi, A.A., and Azen, S.P. (1972). *Statistical analysis: A computer-oriented approach*. New York: Academic Press.

Agretsi, A. (1984). *Analysis of ordinal categorical data*. New York: Wiley-Interscience.

Anderberg, M.R. (1973). *Cluster analysis for applications*. New York: Academic Press.

Anderson, E. (1935). The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society, 59*, 2-5.

Ansfield, F., et.al. (1977). A phase III study comparing the clinical utility of four regimens of 5-fluorouracil. *Cancer, 39*, 34-40.

Bailey, B.J.R. (1980). Large sample simultaneous confidence intervals for the multinomial probabilities based on transformations of the cell frequencies. *Technometrics, 22*, 583-589.

Barnett, V., and Lewis, T. (1978). *Outliers in statistical data*. New York: John Wiley & Sons, Inc.

Bartlett, M.S. (1947). Multivariate analysis. *Journal of the Royal Statistical Society, Series B, 9*, 176-197.

Bates, D.M., and Watts, D.G. (1988). *Nonlinear regression analysis and its applications*. New York: John Wiley & Sons, Inc.

Beale, E.M.L., and Little, R.J.A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society, 37*, Series B, 129-145.

Bebbington, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics, 24*, 136.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons, Inc.

Bendel, R.B., and Afifi, A.A. (1977). Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical Association, 72*, 46-53.

© 1989, SYSTAT, Inc.

Bennett, R.M., and Desmarais, R.N. (1976). Curve fitting of aeroelastic transient response data with exponential functions. In *Flutter testing techniques*. Washington: NASA.

Berger, J.O. (1985). *Statistical decision theory and Bayesian analysis, 2nd Ed.* New York: Springer-Verlag.

Bernhardson, C.S. (1975). Type I error rates when multiple comparison procedures follow a significant *F* test of ANOVA. *Biometrics, 31,* 229-232.

Bhat, M.V., and Haupt, A. (1976). An efficient clustering algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, 6,* 61-64.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Bloomfield, P. (1976). *Fourier analysis of time series: An introduction*. New York: John Wiley & Sons, Inc.

Bock, R.D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Borg, I. (1981). *Anwendungsorientierte multidimensionale skalierung*. Berlin: Springer-Verlag.

Borg, I. (Ed.) (1981). *Multidimensional data representations: When and why?* Ann Arbor: Mathesis Press.

Box, G.E.P., and Cox, D.R. (1964). Analysis of transformations. *Journal of the Royal Statistical Society, Series B, 26,* 211-252.

Box, G.E.P., and Jenkins, G.M. (1976). *Time series analysis: Forecasting and control*. Revised Edition. Oakland, CA: Holden-Day, Inc.

Box, G.E.P., and Muller, M.E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics, 29,* 610-611.

Box, G.E.P., and Tiao, G.C. (1973). *Bayesian inference in statistical analysis*. Reading, Mass.: Addison-Wesley.

Brigham, E.O. (1974). *The fast Fourier transform*. New York: Prentice-Hall.

© 1989, SYSTAT, Inc.

Brownlee, K.A. (1965). *Statistical theory and methodology in science and engineering*. New York: John Wiley & Sons, Inc.

Campbell, D.T., and Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Carmer, S.G., and Swanson, M.R. (1973). An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association, 68,* 66-74.

Carroll, J.D., and Arabie, P. (1980). Multidimensional scaling. In M.R. Rosenzweig and L.W. Porter (Eds.), *Annual Review of Psychology, 31,* 607-649.

Chambers, J.M. (1977). *Computational methods for data analysis*. New York: John Wiley & Sons, Inc.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983). *Graphical methods for data analysis*. Belmont, CA: Wadsworth International Group.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74,* 829-836.

Cleveland, W.S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician, 35,* 54.

Cohen, J., and Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences, 2nd Ed.* Hillsdale, N.J.: Lawrence Erlbaum.

Conover, W.J. (1980). *Practical nonparametric statistics, 2nd. Ed.* New York: John Wiley & Sons, Inc.

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics, 19,* 15-18.

Cooley, J.W., and Tukey, J.W. (1965). An algorithm for the machine computation of complex Fourier series. *Mathematical Computation, 19,* 297-301.

© 1989, SYSTAT, Inc.

Cormack, R.M. (1971). A review of classification. *Journal of the Royal Statistical Society, Series A, 134*, 321-367.

Cornell, J.A. (1985). Mixture experiments. In Kotz, S., and Johnson, N.L. (Eds.), *Encyclopedia of statistical sciences, volume 5*. 569-579. New York: John Wiley & Sons, Inc.

Cox, D.R. (1970). *The analysis of binary data*. New York: Halsted Press.

Dallal, G.E. and Wilkinson, L. (1986). An analytic approximation to the distribution of Lilliefor's test statistic for normality. *American Statistician, 40*, 294-296.

Daniel, C. (1959). The use of half-normal plots in interpreting factorial two level experiments. *Technometrics, 1*, 311-341.

Daniel, C. (1960). Locating outliers in factorial experiments. *Technometrics, 2*, 149-156.

Daniel, C., and Wood, F.S. (1971). *Fitting equations to data*. New York: John Wiley & Sons, Inc.

Davis, D.J. (1977). An analysis of some failure data. *Journal of the American Statistical Association*, 72, 113-150.

Dempster, A.P. (1969). *Elements of continuous multivariate analysis*. San Francisco: Addison-Wesley.

Dennis, J.E. Jr., and Schnabel, R.B. (1983). *Numerical methods for unconstrained optimization and nonlinear equations*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Diamond, W.J. (1981). *Practical experiment designs for engineers and scientists*. Belmont, CA: Lifetime Learning Publications.

Dixon, W.D. and Massey, F.J. Jr. (1969). *Introduction to statistical analysis, 3rd. ed.* New York: McGraw-Hill.

Dixon, W.J. (Ed.) (1981). *BMDP Statistical Software: 1981*. Berkeley: University of California Press.

Doane, D.P. (1976). Aesthetic frequency classifications. T*he American Statistician, 30*, 181-183.

© 1989, SYSTAT, Inc.

Draper, N.R., and Smith, H. (1981). *Applied regression analysis, 2nd Ed.* New York: John Wiley & Sons, Inc.

Einot, I., and Gabriel, K.R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association, 70*, 574-583.

Ekman, G. (1954). Dimensions of color vision. *Journal of Psychology, 38*, 467-474.

Everitt, B.S. (1980). *Cluster analysis, 2nd. Ed.* London: Heineman Education Books, Ltd.

Feingold, M., and Korsog, P.E., (1986). The correlation and dependence between two *F* statistics with the same denominator. *The American Statistician, 40*, 218-220.

Fienberg, S.E. (1979). Graphical methods in statistics. *The American Statistician, 33*, 165-178.

Fienberg, S.E. (1980). *The analysis of cross-classified categorical data, 2nd. Ed.* Cambridge, Mass: MIT Press.

Fisher, L., and Van Ness, J.W. (1971). Admissible clustering procedures. *Biometrika, 58*, 91-104.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Phbilosophical Transactions of the Royal Society of London (A), 222*, 309-368.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179-188.

Flack. V.F., and Chang P.C. (1987). Frequency of selecting noise variables in subset regression analysis: A simulation study. *The American Statistician, 41*, 84-86.

Fletcher, R. (1972). FORTRAN subroutines for minimization by quasi-Newton methods. *AERE* R. 7125.

Freedman, D. A. (1983). A Note on Screening Regression Equations. *The American Statistician*, 37, 152-155.

Freedman, D., Pisani, R., and Purves, R. (1980). *Statistics*. New York: W. W. Norton & Co.

© 1989, SYSTAT, Inc.

Gardner, E.S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, 4, 1-28.

Gentleman, W.M., and Sande, G. (1966). Fast Fourier transforms - for fun and profit. Fall Joint Computer Conference, *AFIPS Conference Proceedings*, *29*, 563-578.

Gilfoil, D.M. (1982). Warming up to computers: A study of cognitive and affective interaction over time. *Proceedings: Human factors in computer systems*. Washington, DC: Association for Computing Machinery.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. New York: John Wiley & Sons, Inc.

Goodman, L.A. (1978). *Analyzing qualitative/categorical data: Log-linear models and latent structure analysis*. Cambridge, MA: Abt Books.

Goodman, L.A., and Kruskal, W.H. (1954). Measures of association for cross-classification. *Journal of the American Statistical Association*, *49*, 732-764.

Gower, J.C. (1985). Measures of similarity, dissimilarity, and distance. In Kotz, S. and Johnson, N.L. *Encyclopedia of statistical sciences, vol. 5*. New York:John Wiley and Sons, Inc.

Gower, J.C. (1967). A comparison of some methods of cluster analysis. *Biometrics*, *23*, 623-637.

Greco, W.R., et.al. (1982). ROSFIT: An enzyme kinetics nonlinear regression curve fitting package for a microcomputer. *Computers and Biomedical Research*, *15*, 39-45.

Green, P.E., and Rao, V.R. (1972). *Applied multidimensional scaling*. New York: Holt, Rinehart, and Winston.

Griffiths, P., and Hill, I.D. (1985). *Applied statistics algorithms*. Chichester: Ellis Horwood Limited.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics*, *25*, 489-504.

Gruvaeus, G., and Wainer, H. (1972). Two additions to hierarchical cluster analysis. *The British Journal of Mathematical and Statistical Psychology*, *25*, 200-206.

© 1989, SYSTAT, Inc.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 139-150.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P.F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*. New York: Free Press.

Guttman, L.A. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, *33*, 469-506.

Haberman, S.J. (1973). Log-linear fit for contingency tables, algorithm AS 51. *Applied Statistics*, *21*, 218-224.

Haberman, S.J. (1978). *Analysis of qualitative data, vol 1: Introductory topics*. New York: Academic Press.

Haberman, S.J. (1982). Analysis of dispersion of multinomial responses. *Journal of the American Statistical Association*, 77, 568-580.

Harman, H.H. (1976). *Modern factor analysis, 3rd. Ed.* Chicago: University of Chicago Press.

Hartigan, J.A. (1975). *Clustering algorithms*. New York: John Wiley & Sons, Inc.

Hartigan, J.A., and Wong, M.A. (1979). A K-means clustering algorithm: Algorithm AS 136. *Applied Statistics*, *28*, 126-130.

Heck, D.L. (1960). Charts of some upper percentage points of the distribution of the largest characteristic root. *Annals of Mathematical Statistics*, *31*, 625-642.

Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (1983). *Understanding robust and exploratory data analysis*. New York: John Wiley & Sons, Inc.

Hocking, R.R. (1983). Developments in linear regression methodology: 1959-82. *Technometrics*, *25*, 219-230.

Hollander, M., and Wolfe, D.A. (1973). *Nonparametric statistical methods*. New York: John Wiley & Sons, Inc.

© 1989, SYSTAT, Inc.

Hubbard, R. and Allen, S.J. (1987). A cautionary note on the use of principal components analysis: Supportive empirical evidence. *Sociological methods and research, 16*, 301-308.

Huber, P.J. (1977). *Robust statistical procedures*. Philadelphia: SIAM.

Jennrich, R.I., and Moore, R.H. (1975). Maximum likelihood estimation by means of nonlinear least squares. In *Proceedings of the statistical computing section*, American Statistical Association, 57-65.

John, P.W.M. (1971). *Statistical design and analysis of experiments*. New York: MacMillan, Inc.

Johnson, S.C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241-254.

Knake, D., and Burke, P.J. (1980). *Log-linear models*. Newbury Park, CA: Sage.

Kritzer, H. (1979). Approaches to the analysis of complex contingency tables: A guide for the perplexed. *Sociological Methods and Research*, 7, 305-329.

Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 1-27.

Kruskal, J.B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29*, 115-129.

Kutner, M.H. (1974). Hypothesis testing in linear models (Eisenhart Model I). *The American Statistician, 28*, 98-100.

Kvålseth, T.O. (1985). Cautionary Note About R^2. *The American Statistician, 39*, 279-285.

Launer, R.L., and Wilkinson, G.N. (1979). *Robustness in statistics*. New York: Academic Press.

Lehmann, E.L. (1975). *Nonparametrics*. San Francisco: Holden-Day, Inc.

Lenstra, J.K. (1974). Clustering a data array and the traveling salesman problem. *Operations Research, 22*, 413-414.

© 1989, SYSTAT, Inc.

Levene, H. (1960). Robust tests for equality of variance. In I. Olkin (Ed.), *Contributions to probability and statistics*. Palo Alto, California: Standard University Press, 278-292.

Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 64*, 399-402.

Ling, R.F. (1973). A computer generated aid for cluster analysis. *Communications of the ACM, 16*, 355-361.

Lingoes, J.C. (1973). *The Guttman-Lingoes nonmetric program series*. Ann Arbor: Mathesis Press.

Lingoes, J.C., and Roskam, E.E. (1973). A mathematical and empirical study of two multidimensional scaling algorithms. *Psychometrika Monograph Supplement, 19*.

Lingoes, J.C., Roskam, E.E., and Borg, I. (Eds.) (1973). *Geometric representations of relational data*. Ann Arbor: Mathesis Press.

Linn, R.L., Centra, J.A., and Tucker, L. (1975). Between, within, and total group factor analyses of student ratings of instruction. *Multivariate Behavioral Research, 10*, 277-288.

Longley, J.W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical Association, 62*, 819-831.

Lovell, M. C. (1983). Data Mining. *The Review of Economics and Statistics, 65*, 1-12.

Lund, R.E., and Lund, J.R. (1983). Probabilities and upper quantiles for the studentized range. *Applied Statistics, 32*, 204-210.

Maindonald, J.H. (1984). *Statistical computation*. New York: John Wiley & Sons, Inc.

Makridakis, W., Wheelwright, S.C. and McGee, U.E. (1983). *Forecasting: Methods and applications, 2nd Ed.* New York: John Wiley & Sons, Inc.

Marascuilo, L.A., and McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Belmont, CA: Wadsworth Publishing.

© 1989, SYSTAT, Inc.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979). *Multivariate analysis.* London: Academic Press.

Marquardt, D.W., and Snee, R.D. (1974). Test Statistics for Mixture Models. *Technometrics, 16,* 533-537.

Marsaglia, G. (1961). *Procedures for generating normal random variables, II.* Boeing Scientific Research Laboratories, Mathematical Note 243.

McCleary, R., and Hay, R.A. Jr. (1980). *Applied time series analysis for the social sciences.* Beverly Hills: Sage Publications.

McGill, R., Tukey, J.W., and Larsen, W.A. (1978). Variations of box plots. *The American Statistician, 32,* 12-16.

McLeod, A.I., and Sales, P.R.H. (1983). An algorithm for approximate likelihood calculation of ARMA and seasonal ARMA models. Algorithm AS 191. *Applied Statistics,* 211-223.

McQuitty, L.L. (1960). Hierarchical syndrome analysis. *Educational and Psychological Measurement, 20,* 293-303.

Metzler, J., and Shepard, R.N. (1974). Transformational studies of the internal representation of three-dimensional objects. In R. Solso (Ed.) *Theories in cognitive psychology: The Loyola Symposium.* Hillsdale, NJ: Lawrence Erlbaum.

Miller, R. (1985). Multiple Comparisons. In Kotz, S., and Johnson, N.L. (Eds.), *Encyclopedia of statistical sciences, volume 5,* 679-689. New York: John Wiley & Sons, Inc.

Milligan, G.W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45,* 325-342.

Morrison, A.S., Black, M.M., Lowe, C.R., MacMahon, B., and Yuasa, S.Y. (1973). Some international differences in histology and survival in breast cancer. *International Journal of Cancer, 11,* 261-267.

Morrison, D.F. (1976). *Multivariate statistical methods.* New York: McGraw-Hill.

Mosteller, F., and Rourke, R.E.K. (1973). *Sturdy statistics.* Reading, MA: Addison-Wesley.

© 1989, SYSTAT, Inc.

Mosteller, F., and Tukey, J.W. (1977). *Data analysis and regression.* Reading, MA: Addison-Wesley, 1977.

Mulaik, S.A. (1972). *The foundations of factor analysis.* New York: McGraw-Hill.

Nelson, C.R. (1973). *Applied time series analysis for managerial forecasting.* San Francisco: Holden-Day, Inc.

Neter, J., and Wassernan, W., and Kutner, M. (1985). *Applied linear statistical models, 2nd Ed.* Homewood, Illinois: Richard E. Irwin, Inc.

O'Neill, R. (1971). Function minimization using a simplex procedure. Algorithm AS 47. *Applied Statistics, 338.*

Pillai, K.C.S. (1960). *Statistical table for tests of multivariate hypotheses.* Manila: The Statistical Center, University of Phillipines.

Pomeranz, J. (1973). Exact cumulative distribution of the Kolmogorov-Smirnov statistic for small samples. *Collected Algorithms from CACM.* Algorithm 487.

Rao, C.R. (1973). *Linear statistical inference and its applications, 2nd Ed.* New York: John Wiley & Sons, Inc.

Rencher, A.C., and Pun, F.C. (1980). Inflation of R-squared in best subset regression. *Technometrics, 22,* 49-54.

Rosenbrock, H. (1960). An automatic method for finding the greatest or least value of a function. *Computer Journal, 3,* 175-184.

Roskam, E.E., and Lingoes, J.C. (1981). MINISSA. In S.S. Schiffman, M.L. Reynolds, and F.W. Young (Eds.), *Introduction to multidimensional scaling: Theory, methods, and applications.* New York: Academic Press, 362-371.

Rothkopf, E.Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology, 53,* 94-101.

Rozeboom, W.W. (1982). The determinacy of common factors in large item domains. *Psychometrika, 47,* 281-295.

© 1989, SYSTAT, Inc.

Rummelhart, D.E. (1970). A multicomponent theory of the perception of briefly exposed visual displays. *Journal of Mathematical Psychology, 7*, 191-218.

Salsburg, D.S. (1985). The religion of statistics as practiced in medical journals. *The American Statistician, 39*, 220-223.

SAS Institute, Inc. (1982). *SAS User's Guide: Statistics, 1982 Edition*. Cary, NC: SAS Institute Inc., 1982.

Scales, L.E. (1985). *Introduction to non-linear optimization*. London: MacMillan Publishers, Ltd.

Schatzoff, M. (1966). Exact distributions of Wilk's likelihood ratio criterion. *Biometrika, 53*, 347-358.

Scheffé, H. (1959). *The analysis of variance*. New York: John Wiley & Sons, Inc.

Schiffman, S.S., Reynolds, M.L., and Young, F.W. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*. New York: Academic Press.

Scott, D.W. (1979). Optimal and data-based histograms. *Biometrika, 66*, 605-610.

Searle, S.R. (1971). *Linear models*. New York: John Wiley & Sons, Inc.

Shepard, R.N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27*, 125-139.

Shepard, R.N. (1963). Analysis of proximities as a study of information processing in man. *Human Factors, 5*, 33-48.

Shepard, R.N., Romney, A.K., and Nerlove, S. (Eds.) (1972). *Multidimensional scaling: Theory and application in the behavioral sciences*. New York: Academic Press.

Shye, S. (Ed.) (1978). *Theory construction and data analysis in the behavioral sciences*. San Francisco: Jossey-Bass, Inc.

Sibuya, M. (1962). On exponential and other random variable generators. *Annals of the Institute of Statistical Mathematics, 13*, 231-237.

© 1989, SYSTAT, Inc.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Simon, A.L. (Ed.) (1969). *Wines of the world*. New York: McGraw-Hill.

Slagel, J.R., Chang, C.L., and Heller, S.R. (1975). A clustering and data reorganizing algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, 5*, 125-128.

Sokal, R.R., and Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin, 38*, 1409-1438.

Sokal, R.R., and Sneath, P.H.A. (1963). *Principles of numerical taxonomy*. San Francisco: W.H. Freeman and Company.

Speed, F.M., and Hocking, R.R. (1976). The use of the R( )- notation with unbalanced data. *The American Statistician, 30*, 30-33.

Speed, F.M., Hocking, R.R., and Hackney, O.P. (1978). Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association, 73*, 105-112.

Spicer, C.C. (1972). Calculation of power sums of deviations about the mean. *Applied Statistics, 21*, 226-227.

Steiger, J.H. (1979). Factor indeterminacy in the 1930's and 1970's: Some interesting parallels. *Psychometrika, 44*, 157- 167.

Sturges, H.A. (1926). The choice of a class interval. *Journal of the American Statistical Association, 21*, 65.

Svalastoga, K. (1959). *Prestige, class and mobility*. London: Heineman Publishing.

Torgerson, W.S. (1958). *Theory and methods of scaling*. New York: John Wiley & Sons, Inc.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Vandaele, W. (1983). *Applied time series and Box-Jenkins models*. New York: Academic Press.

© 1989, SYSTAT, Inc.

Velicer, W.F. (1977). An empirical comparison of the similarity of principal component, image, and factor patterns. *Multivariate Behavioral Research, 12,* 3-22.

Velicer, W.F., Peacock, A.C., and Jackson, D.N. (1982). A comparison of component and factor patterns: A Monte Carlo approach. *Multivariate Behavioral Research, 17,* 371-388.

Velleman, P.F., and Hoaglin D.C. (1981). *Applications, basics, and computing of exploratory data analysis.* Belmont, CA: Duxbury Press.

Velleman, P.F., and Welsch, R.E. (1981). Efficient computing of regression diagnostics. *The American Statistician, 35,* 234- 242.

Wainer, H., and Schacht, S. (1978). Gapping. *Psychometrika, 43,* 203-212.

Wainer, H., and Thissen, D. (1981). Graphical data analysis. *Annual Review of Psychology, 32,* 191-241.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58,* 236-244.

Weber, R. (1982). *EDP auditing: Conceptual foundations and practices.* New York: McGraw-Hill.

Wichman, B.A., and Hill, I.D. (1982). An efficient and portable pseudo-random number generator. Algorithm AS 183. *Applied Statistics, 311,* 188-190.

Wilkinson, J.H., and Reinsch, C. (Eds.). (1971). Linear Algebra, Vol. 2, *Handbook for automatic computation.* New York: Springer Verlag.

Wilkinson, L. (1973). An assessment of the dimensionality of Moos' social climate scale. *American Journal of Community Psychology, 1,* 342-350.

Wilkinson, L. (1975a). REGM: A multivariate general linear hypothesis program for least squares analysis of multivariate data. *Behavior Research Methods and Instrumentation,* 7, 485- 486.

Wilkinson, L. (1975b). Response variable hypotheses in the multivariate analysis of variance. *Psychological Bulletin, 82,* 408-412.

© 1989, SYSTAT, Inc.

Appendix IV: References

Wilkinson, L. (1977). Confirmatory rotation of MANOVA canonical variates. *Multivariate Behavioral Research, 12,* 487-494.

Wilkinson, L. (1978). Permuting a matrix to a simple structure. *Proceedings of the American Statistical Association.*

Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin, 86,* 168-174.

Wilkinson, L. (1980). REGM: A multivariate general linear hypothesis program. *American Statistician, 34,* 182.

Wilkinson, L., and Dallal, G.E. (1977). Accuracy of sample moments calculations among widely used statistical programs. *The American Statistician, 31,* 128-131.

Wilkinson, L., and Dallal, G.E. (1982). Tests of significance in forward selection regression with an F-to-enter stopping rule. *Technometrics, 24,* 25-28.

Winer, B.J. (1971). *Statistical principles in experimental design, 2nd. Ed.* New York: McGraw-Hill.

© 1989, SYSTAT, Inc.

© 1989, SYSTAT, Inc.

Index

[A matrix] (see **MGLH/Test...**)
ABS function (see *Getting Started* and *DATA*)
Absolute value function (see *Getting Started* and *DATA*)
ACF plot... (see **Series**)
ACS function (see *Getting Started* and *DATA*)
Active work area, 406, 412–14, 417–19, 431–33
Adding variables (see *Getting Started* and *DATA*)
Addition (+) (see *Getting Started* and *DATA*)
Additional line, continuing commands on an (see *Getting Started* and *DATA*)
[Additive] (see **Series/Exponential.../Seasonal...**)
Adjusted cell means, 148, 187, 236–37
  Saving, 181, 236
Adjusted squared multiple R, 159, 228
Advanced applications of SYSTAT BASIC (see *DATA*)
Aggregate statistics, 454
  Saving, 454
  Merging with original file, 455
Alienation, coefficient of, 102, 107, 110
[All] (see **MGLH/Test..., Stats/Statistics...**)
Alpha to enter (see **MGLH**, stepwise regression)
Alpha to remove (see **MGLH**, stepwise regression)
Alternative hypothesis, 15, 125
Amalgamation clustering (Linkage clustering), 30–36, 44
Ampersand (&), 153–55, 189–93, 289
Analysis of covariance (ANCOVA), 234–37, 274–77
Alternative to repeated measures, 277
  Homogeneity of slopes, 235
  Multiple covariates, 237
  Repeated measures, 274–77
  Saving adjusted cell means, 237
Analysis of variance (ANOVA), 117–18, 136, 139, 180–280
  Beta coefficients, 200
  Dummy variables, 185, 187–88, 196–98, 204–06, 208, 214
  Fractionalized factorial, 221–23
  Incomplete block, 220–21
  Latin square, 230–31
  Mixed, 223–24

© 1989, SYSTAT, Inc.

Multivariate, 183, 187, 238–44, 254
Multi-group, 138–139
Multiple comparisons, 200–02, 215
Nested, 225–27
One-way, 195–202
Post-hoc tests, 181, 197–99
Printing effects, 200
Profile analysis, 245–50
Randomized block, 219
Repeated measures, 254–56
Split plot, 227–230
Two-way, 117, 202–06, 238, 365
Unbalanced, 181–187, 216–18
ANCOVA (see Analysis of covariance)
AND (see *Getting Started* and *DATA*)
ANOVA (see Analysis of variance)
Appending files vertically (see *DATA*)
[AR parameters] (see **Series/ARIMA...**)
Arc hyperbolic tangent function (see *Getting Started* and *DATA*)
Arccosine function (see *Getting Started* and *DATA*)
Arcsine function (see *Getting Started* and *DATA*)
Arctangent function (see *Getting Started* and *DATA*)
ARIMA... (see **Series**)
Arithmetic functions (see *Getting Started* and *DATA*)
Array variables (see *Getting Started* and *DATA*)
Artificial cases (see *Getting Started* and *DATA*)
ASCII files (see *Getting Started* and *DATA*)
ASN function (see *Getting Started* and *DATA*)
Asterisk (*), 155, 190–01, 193, 289
Asymmetric Lambda, 486
Asymptotic errors (see **Nonlin**)
ATH function (see *Getting Started* and *DATA*)
ATN function (see *Getting Started* and *DATA*)
Autocorrelation, 150, 188, 332, 385–86, 390, 419
Autocorrelation function plots (see **Series/ACF plot...**)
AutoRegressive Integrated Moving Average (ARIMA) modeling, 376, 383–93, 397–98, 404, 414–15, 419, 425–33
Autoregressive modeling, 386–87, 426
Average linkage clustering, 31
[Average] (see **Cluster/Join...**)
AVG function (see *Getting Started* and *DATA*)

© 1989, SYSTAT, Inc.

[Backcast] (see **Series/ARIMA...**)
Backup disks (see *Getting Started*)
Badness of fit (chi-square), 482
Bartlett chi-square, 62, 67, 244, 468
[Base] (see **Series/Transform...**)
BASIC (see SYSTAT BASIC)
Batch files (see Command files)
BCF function (see *Getting Started* and *DATA*)
Beginning of file variable (BOF) (see *DATA*)
Beginning of group variable (BOG) (see *DATA*)
Beta coefficients, 200
Between subjects effects, 257–58, 260–61, 265–66, 270
Bias, 10, 133
BIF function (see *Getting Started* and *DATA*)
Binary similarity coefficients, 59
BOF variable (see *DATA*)
BOG variable (see *DATA*)
Bonferonni-adjusted probabilities, 67–68
Bonferroni procedures, 201, 216, 218
[Bonferonni] (see **Corr/Pearson...**)
Box's F approximation, 488
Box-Jenkins approach, 375
BRN function (see *Getting Started* and *DATA*)
By groups ..., 19, 209, 236, 451, 454, 460, 485-87 (also see *Getting Started* and *DATA*)

[C matrix] (see **MGLH/Test...**)
[Cancel] (see **Cluster/Kmeans...**)
Canonical coefficients, 154, 190, 240, 290–91, 296
Canonical correlation analysis, 157, 283, 289–90, 304–05
Canonical loadings, 157, 244
Canonical variable plot, 298
Caret (^) (see *Getting Started* and *DATA*)
Case number (see *Getting Started* and *DATA*)
Case plot... (see **Series**)
Case weighting (see *Getting Started*)
Cases, maximum number of (see *Getting Started*)
Categorical data, 118, 471
[Categories] (see **MGLH/General linear model...**)
Caution messages (see *Getting Started*)
CCF... (see **Series**)

© 1989, SYSTAT, Inc.

CDF (see ZCF) (see *Getting Started* and *DATA*)
Cell means, 148, 187, 189, 191, 208–09, 216, 236–37, 250–52, 451
[Centroid] (see **Cluster/Join...**)
Changeover designs, 232–33
Changing variables (see *Getting Started* and *DATA*)
Changing variable names (see *Getting Started* and *DATA*)
Character functions (see *Getting Started* and *DATA*)
Character values (see *Getting Started* and *DATA*)
Character variables (see *Getting Started* and *DATA*)
Chi-square approximation, 476
Chi-square statistic, 244, 476, 482–83, 489, 498–99
Chi-square test, 62, 67, 353, 363, 476, 479–82
[Chi-square] (see **Npar/Kolmogorov-Smirnov...**)
City... (see **Corr**)
City block distances, 21, 58–59, 63, 102
[Class variables] (see **Tables/Loglinear models...**)
Classification table, 297
Clear...(see **Series**)
[Clear] (see *Getting Started*)
Cluster..., 18–46
  Algorithms, 36
  Clustering columns (cases), 24
  Clustering rows (variables), 23
  Clustering rows and columns, 24–25
  Hierarchical methods (Tree methods), 20–21, 23–25, 27–33, 37–39, 44–46
  Join..., 19–20, 27–33, 36, 38–42
    [Average], 31
    [Centroid], 38
    [Columns], 28
    [Complete], 30
    [Euclidean], 28
    [Gamma], 28
    [Label cases with], 30
    [Matrix], 29
    [Median], 31
    [Pearson], 28
    [Percentage], 29
    [Rows], 28
    [Save identifiers], 31
    [Single], 30
    [Ward], 31

© 1989, SYSTAT, Inc.

  Choosing distance metrics, 28–29
  Choosing linkage (amalgamation) methods, 30
  Direct display of a matrix, 33
  Joining columns, 29
  Joining rows, 29
  Joining a matrix, 29
  Missing values, 29–30, 33
Kmeans..., 25–26, 34–35, 42–44
  Adjusting number of groups formed, 35
  Specifying the number of iterations, 34
  [Groups], 35
  [Iterations], 34
  [Label], 36
  [Save identifiers], 33
Labeling a tree, 30, 32
Linkage methods, 30–31
"Max" method, 30
Memory limitations, 36
"Min" method, 36
Missing data, 36
Passing matrices from **Corr**, 29
Single linkage clustering, 21, 24, 30
Tree methods (Hierarchical methods), 20–21, 23–25, 27–33, 37–39, 44–46
Two-way clustering, 42
Ward's method, 31
Cluster centroid, 30
CODE (see *DATA*)
Coefficient of determination (see squared multiple R)
Coefficient of variation, 453
Cohen's kappa, 481, 485
COLD commands, 514–515
[Column pct] (see **Tables/Tabulate...**)
[Columns] (see **Cluster/Join...**)
Command files, 508, 511
Command reference, 506–53
Command window, 63, 168, 296, 507–08 (also see *DATA*)
Commands, using, 506–53 (also see *DATA*)
Comments on output (also see *Getting Started*)
Common factor analysis, 79–83, 87
Common factor model... (see **Factor**)
Communality, 78

© 1989, SYSTAT, Inc.

[Complete] (see **Cluster/Join...**)
[Component or factor loadings] (see **Factor/Saving files**)
[Component score coefficients] (see **Factor/Saving files**)
[Component scores] (see **Factor/Saving files**)
[Component] (see **Series/Exponential...**)
Compound symmetry, 278
Condensing records (see *DATA*)
Condition indices, 157, 163
Conditional transformations (see IF...THEN command)
[Confi] (see **Tables/Tabulate...**)
Confidence interval, 13–16, 353, 476, 481, 485, 493
  Asymptotic, 433
  Simultaneous, 157
Configuration... (see **MDS**)
Confirmatory components, 75
Confirmatory models, 79, 376
Confirmatory scaling, 104
[Constant] (see **Series/ARIMA...**)
Contingency, 481, 485
Continuing on an additional line, 518
Continuity correction, 363, 493, 497
Contour plots, 349
[Contrast] (see **MGLH/Test...**)
Contrasts, 150, 157, 191–93, 200, 208, 248–50
Convergence, controlling, 86, 101, 316–17, 319, 321, 405
Cook's D, 130, 149, 159–60, 187
Corr, 48–58
  Algorithms, 64
  City..., 21, 58–59, 63–64
  Covariance..., 58–59, 152
  Dichotomous..., 58–59
    [S2], 58–59
    [S3], 58–59
    [S4], 58–59, 58–59
    [S5], 60
    [S6], 58–59
  Euclidean..., 60
  Gamma..., 21, 60
  Missing data, 56–57, 64
    [Listwise] deletion, 56–57
    [Pairwise] deletion, 56–57

© 1989, SYSTAT, Inc.

  Mu2..., 21, 61
  Pearson..., 21, 51–52, 61–62, 65
    [Bonferonni], 62, 67–68
    [Probability], 62, 65–67
    Probabilities for Pearson correlation coefficients, 62
    [Save file], 63
  Spearman...,21, 62, 64
  SSCP..., 62, 64
  Tau..., 21, 58, 63, 64
  Triangular matrices, 140, 152
Correlation matrices, 23, 25, 52, 55–56, 58, 62–64, 67, 71–71, 81, 83, 86–87, 94, 96, 100, 155, 157, 291, 301–02
  As input, 150–52, 285–86
Correlation matrix of regression coefficients, 157, 164
Correlation metric, joining with, 38, 39
[Correlation] (see **Factor**)
COS function (see *Getting Started* and *DATA*)
Cosine function (see *Getting Started* and *DATA*)
Covariance... (see **Corr**)
Covariance matrix, 72, 76, 78–79, 83, 86, 291, 302
  As input, 150–52, 285–86,
[Covariance] (see **Factor**)
[Covariates] (see **MGLH/Fully factorial (M)ANOVA...**)
Cramer's V, 481, 485
Creating a data file (see *Getting Started* and *DATA*)
Creating an ASCII file from a SYSTAT file (see *Getting Started* and *DATA*)
[Criterion] (see **Tables/Loglinear models...**)
Critical level, 15, 125
Cross products, 56, 62, 243, 280, 291
Cross tabulations (see **Tables**)
Cross validation, 90, 134, 149, 177, 287, 299
Cross-correlation function plots (CCF), 404, 419, 422–24
Crossover designs, 232–34
Cumulative distribution function, 368, 371
Current options (see *Getting Started*)
[Cut] (see **Npar/Wald-Wolfowitz runs...**)
Cutpoints, 33, 473
[CV] (see **Stats/ Statistics...**)

© 1989, SYSTAT, Inc.

[D matrix] (see **MGLH/Test...**)
D-statistic (see Cook's D)
DATA (see *DATA*)
Data Editor (see *Getting Started*)
Data files (see *Getting Started* and *DATA*)
  ASCII (see ASCII files)
  SYSTAT (see SYSTAT data files)
Data types (see *Getting Started* and *DATA*)
Data window (see *Getting Started* and *DATA*)
Decimal places
  Shown in output, 62, 469, 486
  Shown in Editor (see *Getting Started*)
Decrement in stress or alienation, 101
[Decrement] (see **MDS/Scale...**)
Degrees of freedom, 151–52, 156, 167, 193, 216, 239, 283, 464, 476, 479, 489
Deleting cases (see *Getting Started* and *DATA*)
[Delta] (see **Tables/Loglinear models...**)
Dendrogram, 21
Density functions (see *Getting Started* and *DATA*)
Dependent variable means, 251
[Dependent] (see **MGLH**)
Derivative-free minimization (Simplex method), 307, 316, 319, 321–22, 324–25
Descending sorts (see *DATA*)
Descriptive statistics, 3–9, 124, 445–46, 449, 453–56
Design matrix, 187, 205, 209
Designed experiments (see Experimental designs)
Deterministic models, 348–49
Detrended values, 417
Detrended series (see Differenced series)
Dichotomous... (see **Corr**)
Dichotomy coefficients, 58–60
  Anderberg's ([S5]), 59–60
  Jacard's ([S3]), 59–60
  Positive matching ([S2]), 59–60
  Simple matching ([S4]), 59–60
  Tanimoto's ([S6]), 59–60
[Diff] (see **MGLH/Test...**)
Difference, 399–400
Difference contrasts (see **MGLH/Test...**)
[Difference] (see **Series/Transform...**)

© 1989, SYSTAT, Inc.

Differenced series, 389, 393, 399–400, 416–19
Digits displayed (see FORMAT command)
Dimension arrays (see *DATA*)
[Dimension] (see **MDS/Configuration...**)
Discrete Fourier transform, 395–96
Discriminant analysis, 117, 139, 283, 289–91, 292–99
Discriminant function, 292–93, 296
Discriminant scores, 296, 299
Displaying joining distances in a tree, 32
Dissimilarity data
  Collecting, 95–96
  Iterations, 97, 101
  MDS plot (see **MDS**)
  Scaling, 97, 100–103
  Shepard diagram, 102
  Stress, 97
Dissimilarity matrix (see also *Getting Started* and *DATA*)
Distance matrix, 36, 44–46
Distance metric, 28–29, 102
Distribution free statistics, 354, 368
Distribution functions, 342, 368, 371
Division (/) (see *Getting Started* and *DATA*)
Domain, 120
Double precision, 63–64, 87, 140, 177, 319, 415, 459, 490
Dropping variables (see *Getting Started* and *DATA*)
Dummy variables, 185, 187–88, 196–98, 204–06, 208, 214
Duncan's multiple range test, 201–02
Durbin-Watson statistic, 150, 188

ECF function (see *Getting Started* and *DATA*)
Ecological data, 168
Editing (see *Getting Started* and *DATA*)
Editing an existing SYSTAT file (see *Getting Started* and *DATA*)
Editing program lines (see *DATA*)
Editor screen (see *Getting Started*)
Effects coding, 138, 147, 185, 187, 284
Effects in analysis of variance, 187–91, 193–94, 196–97, 200, 208–09
[Effects] (see **MGLH/Test...**)
EIF function (see *Getting Started* and *DATA*)
Eigenvalue, 85, 157, 163–64, 244

© 1989, SYSTAT, Inc.

[Eigenvalue] (see **Factor**)
Eigenvalues of unit scaled X'X, 157, 163
End of file variable (EOF) (see *DATA*)
End of group variable (EOG) (see *DATA*)
[Enter] (see **MGLH/General linear model...**)
Entering data (see *Getting Started* and *DATA*)
Entering editor (see *Getting Started*)
Entering variable names (see *Getting Started* and *DATA*)
EOF variable, 315, 349 (also see *DATA*)
EOG variable, 349 (also see *DATA*)
Equal to (=) (see *Getting Started* and *DATA*)
Equally spaced polynomials, 194, 254
[Equamax] (see **Factor**)
Erase program lines (see *DATA*)
ERN function (see *Getting Started* and *DATA*)
Error matrix, Factoring, 290
Error messages (see *Getting Started* and *DATA*)
Error sum of products matrix, 157, 291
Error sum of squares, 166, 173, 302
Error term for ANOVA, 193, 206, 223–24, 227–29, 240, 277, 280
ESTIMATE, 127, 148–49, 160, 164, 187–88, 209, 233, 319, 324
Estimates of effects, 216–17, 240
Estimating a model, 12–13, 385
[Estimation method] (see **Nonlin/Model...**)
Euclidean... (see **Corr**)
Euclidean distance
  Computing (see **Corr**)
  Joining with (see **Cluster**)
[Euclidean] (see **Cluster/Join...**)
Evaluation of expressions (see *Getting Started* and *DATA*)
Exclusive clusters, 20
EXP function (see *Getting Started* and *DATA*)
Experimental designs, 188, 225
Exploratory data analysis, 3–4, 75, 133, 376, 383
Exponential... (see **Series/Exponential...**)
Exponential distribution functions (see *Getting Started* and *DATA*)
Exponential function (see *Getting Started* and *DATA*)
Exponential smoothing 394–95, 408–10, 434–37
Exponentiation (^) (see *Getting Started* and *DATA*)
Expressions (see *Getting Started* and *DATA*)
Extrapolating, 120, 385, 392

© 1989, SYSTAT, Inc.

F distribution (see also *Getting Started* and *DATA*)
F ratio, 25, 44, 126, 137, 160, 196
F statistic, 167, 198, 201, 243–44, 254, 274, 278
F test, 177, 189, 191, 211, 215–217, 243, 247, 249
F-value, 34, 44, 175, 201, 218
Factor, 70–91
  Algorithms, 87
  Choosing number of factors to retain, 85
  Choosing number of components to retain, 85
  Common factor model..., 86
    [Correlation], 84
    [Covariance], 84
    [Eigenvalue], 157, 163
    [Iterations], 86
    [Labels], 86
    [Number of factors], 85
    [Pairwise deletion], 84
    [Plot factors], 85
    [Save file], 86
    [Sort loadings], 85
    [Tolerance], 86
  Component coefficients, 76
  Component loadings, 76
  Convergence criterion, 86
  Exploratory vs. confirmatory components, 75
  Factor analysis, 78, 79
  Iterated principal axis (**Common**) factor analysis, 71, 83
  "Maximum likelihood" factor analysis, 79, 82
  Missing data, 87
  Missing values, 84
  Plots of loadings, 85
  Principal components..., 83–86
    [Correlation], 84
    [Covariance], 84
    [Eigenvalue], 157, 163
    [Num of factors], 84
    [Plot factors], 85
    [Save file], 85
    [Sort loadings], 85
  Principal components analysis versus factor analysis, 79–82
  Ridge regression, 79

© 1989, SYSTAT, Inc.

Rotations, 76
[Equamax], 77
[Quartimax], 77
[Varimax], 77, 85
Saving files, 55
[Coefficients], 85
[Loadings], 85
[Scores], 85
Signs of loadings, 77
Factor analysis versus principal components analysis, 79–82
[Factor] (see **MGLH/Test...**)
Factoring error matrix, 290
[Factors] (see **MGLH/Fully factorial (M)ANOVA...**)
Fast Fourier transform, 395, 414
FCF function (see *Getting Started* and *DATA*)
Field formats (see *DATA*)
FIF function (see *Getting Started* and *DATA*)
File comments (see *Getting Started* and *DATA*)
File manipulation (see *Getting Started* and *DATA*)
Filenames (see *Getting Started* and *DATA*)
Filtering series, 375, 377, 380–82, 395–96, 407
Finding specific values (see *Getting Started*)
[First period] (see **Series/ID...**)
Fisher and Yates approach, 218
Fisher discriminant functions, 296
Fisher's exact test, 480, 486
Fitted values, 318, 344, 482, 489–90
[Fitted] (see **Tables/Loglinear models...**)
Fixed-format input (see *DATA*)
Floppy disk installation (see *Getting Started*)
Force variables into equation (see **MGLH**, stepwise regression)
[Force] (see **MGLH/General linear model...**)
[Forecast] (see **Series/ARIMA.../Exponential...**)
Forecasting, 383–93, 405–06, 410, 415, 432–33
Format of numbers (see *Getting Started* and *DATA*)
Formatted input (see Fixed-format input)
FOR...NEXT loop (see *DATA*)
Fourier... (see **Series**)
Fourier analysis, 376, 393, 395–400, 412–13, 439–42
Fractional factorial designs, 221–23
Free-format input (see *DATA*)
Frequency domain, 376

© 1989, SYSTAT, Inc.

Frequency tables (see **Tables**)
[Frequency] (see **Tables/Tabulate...**)
Friedman... (see **Npar**)
Full screen data editor (see *Getting Started*)
Fully crossed designs, 181, 312
Fully factorial (M)ANOVA... (see **MGLH**)
Fully saturated log-linear model, 481–82, 488
Function plots (see *SYGRAPH*)
Functions (see *Getting Started* and *DATA*)

Gamma... (see **Corr**)
Gamma coefficients, 58, 60, 471
[Gamma] (see **Cluster/Join...**)
Gapping, 33
GCF function (see *Getting Started* and *DATA*)
General features of SYSTAT (see *Getting Started*)
General linear model... (see **MGLH**)
GIF function (see *Getting Started* and *DATA*)
Goodman-Kruskal gamma correlation coefficient, 58, 60, 471
Goodness-of-fit (chi-square), 97, 176
GOTO (see *DATA*)
Gramian matrices, 53
Grand mean, 193, 205, 216, 237, 252
Graphical displays (see *SYGRAPH*)
Greater than (>) (see *Getting Started* and *DATA*)
Greater than or equal to (>=) (see *Getting Started* and *DATA*)
GRN function (see *Getting Started* and *DATA*)
Group classification constants, 296
Group classification function coefficients, 296
[Group] (see **Stats/Statistics...**)
Grouped data (see *Getting Started* and *DATA*)
Grouping variables (see *Getting Started* and *DATA*)
[Groups](see **Cluster/Kmeans...**)
Guttman and Kruskal scaling (see **MDS/Scale...**)
Guttman method, 102, 105, 106–10
Guttman mu coefficients, 52, 58, 61, 353
Guttman simplex, 41–42
Guttman Smallest Space Analysis, 100
Guttman/Lingoes coefficient of alienation, 102, 107, 110
[Guttman] (see **MDS/Scale...**)

© 1989, SYSTAT, Inc.

Hanning weights, 380
Hard disk installation (see *Getting Started*)
Hessian matrix, 319–20, 323–24, 415
Hierarchical clustering (see **Cluster**)
HOLD (see *DATA*)
Homogeneity of slopes assumption, 235–36, 276–77
Homogeneity of variance, 137, 170, 197, 209, 455, 468
HOT commands, 514, 515
Hotelling's T^2, 237–38
  One-sample test, 237
  Two-sample test, 238
Hotelling-Lawley trace, 243
Householder tridiagonalization, 37, 140
Hyperbolic arctangent (see *Getting Started* and *DATA*)
[Hypoth] (see **MGLH/Test...**)
Hypothesis, definition, 9
Hypothesis matrix
  Factoring, 290
Hypothesis sum of product matrix, 157, 243, 291
Hypothesis sum of squares, 166, 171, 243, 278
Hypothesis testing, 14–16, 124–26, 152–56, 165–68, 188–94, 240–43, 287–91, Chapter 6 *passim*
  Non-zero null hypothesis, 168
  One regression coefficient, 165–66
  Selecting effects, 190–91, 289
  Two regression coefficients, 166–67

ID... (see **Series**)
IF...THEN command (see *DATA*)
IF...THEN...ELSE format (see *DATA*)
Incomplete blocks, 220–21
[Indep'dent] (see **MGLH/General linear model...**)
[Independent] (see **MGLH/Regression..., Stats/t-test...**)
[Index] (see **Series/Transform...**)
Inequalities (see *Getting Started* and *DATA*)
Inference, 3, 9–17, 122–24, 475
Inferential statistics, 124
Initial configuration, 105–07
Input formats (see *DATA*)
INT function (see *Getting Started* and *DATA*)
Integer truncation function (see *Getting Started* and *DATA*)

© 1989, SYSTAT, Inc.

Interaction term, 147, 155, 185–86, 191, 203, 273, 277, 284, 286, 289, 481, 503
[Interactions] (see **MGLH/General linear model...**)
Interpolating, 120
Interquartile range (Midrange), 447
Inverse contrast, 241
Inverse distribution (see *Getting Started* and *DATA*)
Inverse distribution functions (see *Getting Started* and *DATA*)
[Iter] (see **Nonlin/Model..., Tables/Loglinear models...**)
Iterated principal axis factoring, 71, 83
Iterations, 34, 86, 97, 101–02, 107–07, 316–17, 319, 321–22, 405, 489
[Iterations] (see **Factor/Common factor model..., MDS/Scale..., Series/ARIMA..., Cluster/Kmeans...**)
Iteratively reweighted least squares, 339–42

Jaccard coefficient, 60
Join... (see **Cluster**)
[Join] (see **Cluster/Join...**)

K-means splitting method of clustering (see **Cluster**)
Kendall coefficient of concordance, 365
Kendall's tau-b, 365
Kinetic models, 345–48
Kmeans... (see **Cluster**)
Kolmogorov-Smirnov (see **Npar**)
Kruskal method, 105–06, 111–15
Kruskal's stress formula, 102
Kruskal-Shepard-Torgerson-Young multidimensional scaling, 100
Kruskal-Wallis... (see **Npar**)
[Kruskal] (see **MDS/Scale...**)
[Kurtosis] (see **Stats/Statistics...**)

LABEL (see *DATA*)
[Label cases with] (see **Cluster/Join...**)
[Label](see **Cluster/Kmeans...**)
Labeling a tree (see **Cluster**)
[Labels] (see **Factor/Common factor model...**)
LAG (see *Getting Started* and *DATA*)
[Lag] (see **Series/ACF.../Case plot.../PACF.../Transform...**)

© 1989, SYSTAT, Inc.

Lagging transformations, 398, 399–400, 402–04, 423, 427–30, 440
Lambda (see Asymmetric lambda or Wilks' lambda)
Latent roots, 302
Latent trait analysis, 75
Latin squares, 230–31
Leakage, 400
Least absolute values (see **Loss function...**)
Less than (<) (see *Getting Started* and *DATA*)
Less than or equal to (<=) (see *Getting Started* and *DATA*)
LET command (see *DATA*)
LEVERAGE (see Leverage)
Leverage, 148–49, 160, 164, 187–88
LGM function (see *Getting Started* and *DATA*)
Likelihood ratio chi-square, 483, 487
[Lilliefors] (see **Npar/Kolmogorov-Smirnov...**)
Limitations of procedures (see *Getting Started*)
Line numbers (see *DATA*)
Linear filter, 375, 378, 407
Linear models, 118–30, 135, 147–52, 184–88, 284–87, 116–305
Linear regression, 142–43, 310–11, 327–28, 384–88
[Linear] (see **MDS/Scale...**, **Series/Exponential...**)
Line, equation, 118–20
Linkage method of clustering (see **Cluster**)
[Linkage] (see **Cluster/Join...**)
[List format] (see **Tables/Tabulate...**)
Listing a program (see *Getting Started* and *DATA*)
Listing a SYSTAT data file (see *Getting Started* and *DATA*)
Listwise deletion of missing data, 56–57, 84
[Listwise] (see **Corr**)
Loadings, 76–77, 81–82, 85–86, 244
LOG function (see *Getting Started* and *DATA*)
Log-linear models, 481–83, 487–91, 499–505
[Log] (see **Series/Transform...**)
Logarithms (see *Getting Started* and *DATA*)
Logical expressions (see *Getting Started* and *DATA*)
Logical operators (see *Getting Started* and *DATA*)
Logical record length (see *DATA*)
Logistics models, 339–42
  Iteratively reweighted least squares, 339–40
  Maximum likelihood estimation, 340–42
Loglinear models... (see **Tables**)
Longley data, 160, 163–64, 167, 173, 176

© 1989, SYSTAT, Inc.

Loops (see *DATA*)
Loss function
  Least absolute values, 312, 318, 322, 328
  Negative log likelihood, 307, 312, 341
  Sum of least squares, 312
[Loss function] (see **MDS/Scale...**)
LOWESS...(see **Series**)

[MA parameters] (see **Series/ARIMA...**)
Mahalanobis distance (DISTANCE), 287, 291, 294
Mann-Whitney U, 353–54, 357, 360
Mantel-Haenszel test, 497–98
Map file transfer (see *SYGRAPH*)
Map files (see *SYGRAPH*)
Matrix, 50
Matrix input (see *Getting Started* and *DATA*)
Matrix of effects, 188
Matrix of linear weights (A matrix), 144, 153, 188, 283, 287
Matrix of pairwise absolute mean differences, 467
[Matrix type] (see **Factor/Common factor model...**)
[Matrix] (see **Cluster/Join...**)
MAX function (see *Getting Started* and *DATA*)
Max method (see **Cluster**)
[Max] (see **Series/Case plot...**)
Maximum absolute correlations, 87
Maximum likelihood (see **Loss function...**)
Maximum likelihood factor analysis (see **Factor**)
Maximum value across variables (MAX function)
[Maximum] (see **Stats/Statistics...**)
McNemar's test, 480
MDS, 92–115
  Algorithms, 105–06
  Assumptions, 94–95
  Choosing a regression function, 102
  Collecting dissimilarity data, 95–96
  Confirmatory scaling, 104
  Controlling convergence, 101
  Configuration..., 100, 102, 103–04
    [Dimension], 103
  Distance metric, 102
  Guttman method, 102, 105, 106–110

© 1989, SYSTAT, Inc.

Kruskal method, 105–06, 11–15
Missing data, 105
Number of dimensions, 101
Scale..., 100–103
[Decrement], 101
[Guttman], 102
[Iterations], 101
[Kruskal], 102
[Linear], 100
[Loss function], 102
[Minimum], 101
[Minkowski constant], 102
[Monotonic], 100
[Save file], 102
[Shepard plots], 102
Mean, 5
Mean square error (MSE), 216, 236, 246, 317, 341, 392
[Mean window] (see **Series/Smooth...**)
[Mean] (see **Series/Transform..., Stats/Statistics...**)
Measures of association for two way tables, 480–81, 483, 486
Measures for 2×2 tables, 481, 485
Median, 5
Median linkage clustering, 31
[Median window] (see **Series/Smooth...**)
[Median] (see **Cluster/Join...**)
Memory limitations of Cluster, 36
Menus (see *Getting Started*)
Merging files horizontally (see DATA)
Meta-analysis, 152, 186
Metric conditions, 94
MGLH, 116–305
Adjusted cell means, 148, 181, 187, 236–237
Algorithms, 140
Between plots effects, 228
Canonical correlation analysis, 244, 304–05
Canonical variable plot, 298
Comparing effects, 191–92
Contrasting effects, 191–92
General linear contrasts, 192–193
Within factors, 191
Correlation matrix as input, 150–52, 285–86
Covariance matrix as input, 150–52, 285–86

© 1989, SYSTAT, Inc.

Cross validation, 90, 134, 149, 177, 287, 299
Default output, 243
Diagnostics, 126–30, 148–49, 187–88
Durbin-Watson statistic, 150, 188
ESTIMATE, 127, 148–49, 160, 164, 187–88, 209, 233
Extended output, 162–64
Factoring, 290–91
Choosing the number of components to rotate, 290
Choosing the type of matrix to factor, 290
First-order autocorrelation, 150, 188
Fully factorial (M)ANOVA..., 182, 183–84
[Covariates], 184
[Dependent], 184
[Factors], 184
[Repeat], 184
[Save residuals], 184
Without constant term, 188
General linear model..., 145, 147–52, 175–78, 182, 184–88, 196, 203–05, 213, 219–23, 226–35, 237–39, 245–47, 251–56, 266–70, 277–78, 284–87, 292–97, 301–02, 304–05
[Cases], 150, 186, 285
[Categories], 147, 185, 203–06, 244, 284
[Dependent], 183, 184–86, 244, 259, 284
[Enter], 150, 177, 185, 285
[Force], 177, 185, 285
[Indep'dent], 145–46, 147, 150, 185, 213, 244, 259, 284, 287
[Interactions], 147, 187, 213, 284
[Mixture model], 150, 173–75, 186, 285
[Next], 147, 185, 284
[Preview], 148, 186, 193, 286
[Remove], 150, 177, 185, 285
[Save file], 148, 186, 196, 236, 287
[Stepwise], 140, 150, 176, 185, 285
Homogeneity of slopes, 235
LEVERAGE, 148–49, 160, 164, 187–88
Missing data, 140, 152, 204, 280, 286
Multiple comparison, 200–02, 215
Multivariate models, 282–308
Canonical correlation, 283, 289–91, 304–05
Discriminant analysis, 283, 289–90, 291, 192–99
Principal components, 283, 289–90, 291, 299–304

© 1989, SYSTAT, Inc.

Previewing your model, 148, 186, 286
Prior probabilities, 154, 190, 289, 298
Profile analysis, 153–55, 189, 191, 194, 245–50, 254, 278, 280, 288
Regression..., 142–78, 184, 235–36
[Dependent], 147
[Independent], 145–47, 150
[Save residuals], 146, 160, 164
Without the constant, 188
Regression coefficients, 125, 135, 137, 140, 143–44, 148–49, 152–53, 155–57, 159, 161–64, 165–67, 182, 187, 194, 196, 214, 240, 246, 283, 286
Repeated measures, 117, 150, 154, 182, 184, 186, 254–80
RESIDUAL, 148–49, 160, 164, 187–88, 233
Saving canonical scores, 296–97
Saving results, 291
SSCP matrix as input, 150–52, 285–86
Selecting error terms, 193, 277
SEPRED, 149, 187
Standardizing canonical coefficients, 290
Single degree of freedom designs, 212–16
Stepwise regression, 134, 150, 176–78, 185, 285
Alpha to enter, 177
Alpha to remove, 177
STUDENT, 127, 149, 160, 164, 187
Submodel selection, 132–34, 142, 177
Backward elimination, 132–33
Bias and variance tradeoff, 133
Dimensionality, 133
Evaluation, 134
Forward selection, 132
Removal criteria, 133
Stepwise, 132
Strategy, 133
Test..., 152–57, 165–68, 171, 177, 188–94, 197–202, 206–09, 216–19, 228–29, 237–40, 223–24, 245–53, 271–74, 278–79, 287–91, 293, 298, 302, 304–05
[All], 155, 165, 191, 289
[A matrix], 155–57, 167, 193, 198–99, 206–09, 240, 246, 283
[C matrix], 153, 155, 189, 191–93, 250, 270–74, 279–80, 288–89
[Contrast], 153, 189, 191–93, 206–07, 288–89
[Corr], 302

© 1989, SYSTAT, Inc.

[D matrix], 153, 155–56, 189, 191–93, 237, 288–89
[Diff], 194, 248, 251, 253, 278–79
[Effects], 153–54, 155, 166–68, 189–93, 216, 271, 283, 288, 289, 294, 299
[Error], 154, 193–94, 218, 223–24, 288, 290, 302
[Factor], 155, 190, 290, 302
[Hypoth], 290, 302
[Poly], 194, 246–47
[Priors], 154, 190, 289, 198
[Profile], 153–55, 189, 191, 194, 245–50, 254, 278, 280, 288
[Save scores and results], 155, 291, 293–94, 296
[SSCP] (see SSCP matrix)
[Stdize], 154, 190, 290, 304
[Sum], 194, 249, 272, 274
Test of simple effects, 181, 193, 208–09
Unbalanced designs, 181, 187, 206, 216–18, 227, 277
Within plots effects, 229–30
Midrange (Interquartile range), 447
MIN function (see *Getting Started* and *DATA*)
Min method (see **Cluster**)
[Min] (see **Series/Case plot...**)
Minimizing an analytic function, 348–49
Minimum value across variables (see MIN function)
[Minimum] (see **MDS/Scale..., Stats/Statistics...**)
[Minkowski constant] (see **MDS/Scale...**)
Minkowski metric, 102
MIS function (see *Getting Started* and *DATA*)
Missing data across variables (see MIS function)
[Missing] (see **Tables/Tabulate...**)
Mixed models, 117, 223–24
[Mixture model] (see **MGLH/General linear model...**)
Model... (see **Nonlin**)
Model constraints, 320–21
Model estimation, 122–24, 173, 316
[Monotonic] (see **MDS/Scale...**)
Monotonicity coefficient (mu), 52, 58, 61, 353
Monte Carlo simulation, 31, 152, 286, 458
Moving average modeling, 388, 393, 404, 426
Moving averages smoothing, 377–80
Mu2... (see **Corr**)
Multi-variable functions (see *Getting Started* and *DATA*)
Multi-way tables, 500–505

© 1989, SYSTAT, Inc.

Multicollinearity, 160, 161, 316
Multidimensional scaling (**MDS**)
Multiple comparisons, 200–02, 215
Multiple correlations, 87, 126, 159, 171, 173, 177, 194, 246
Multiple R, 159, 228
Multiple R squared, 159, 228
Multiple regression, 130–35, 160–64
Multiplication (*) (see *Getting Started* and *DATA*)
[Multiplicative] (see **Series/Exponential.../Seasonal...**)
Multivariate analysis of variance (MANOVA), 183, 187, 238–44, 254
Multivariate General Linear Hypothesis (see **MGLH**)
Multivariate general linear model, 135
Multivariate models, 282–305
  Canonical correlation, 283, 289–91, 304–05
  Discriminant analysis, 283, 289–90, 291, 192–99
  Principal components, 283, 289–90, 291, 299–304
Multivariate profile analysis, 245–50
  Difference contrasts, 248–49
  Special contrasts, 250
  Summing effects, 249
  Unevenly spaced polynomials, 247–48
Multivariate regression, 134–35, 152, 157, 244, 285–86
Multivariate test statistics, 243, 302

[N] (see **Stats/Statistics...**)
Naming data files (see *Getting Started* and *DATA*)
Naming variables (see *Getting Started* and *DATA*)
Natural logarithm function (see *Getting Started* and *DATA*)
Negative log likelihood (see **Nonlin/Loss function...**)
Nested designs, 117, 225–27
Nested loops (see *DATA*)
Nested sort (see *Getting Started* and *DATA*)
Newman-Keuls test, 458
[Next] (see **MGLH/General linear model...**)
Nominal data, 354–55
Non-hierarchical clustering (see **Cluster/Kmeans...**)
Nonlin, 175–76, 306–350
  Algorithms, 323
  Asymptotic standard errors, 319
  Convergence, 316–17
  Deterministic models, 348–49

© 1989, SYSTAT, Inc.

Hessian matrix, 319–20, 323–24
Initial values for parameters
Kinetic models, 345–48
Linear regression, 310–311, 327–28, 384–88
Loss function..., 311–312, 317
  Contouring, 349–50
  Least absolute values, 312, 318, 322, 328
  Maximum likelihood, 312–17
  Negative log likelihood, 307, 312, 341
  Sum of least squares, 312
Mean estimation, 325
Median estimation, 326–27
Minimizing an analytic function, 348–49
Missing data, 323
Model..., 313–17
  Estimation method, 315
  [Iter], 316
  [Num of factors], 315
  [Quasi-Newton], 315
  [Print], 317
  [Rescale MSE], 316
  [Save file], 317
  [Simplex], 315
  [Start values], 316
  [Tolerance], 316
Model constraints, 320–21
Parameter bounds, 320–21
Parameters intercorrelations, 319–20
Piecewise regression, 333–337
Probit models, 342–45
Problems, 321–22
Random model, 349
Resume..., 319
Warning messages, 320
Nonlinear estimation (see **Nonlin**)
Nonlinear filters, 375, 380–81
Nonorthogonal designs, 206, 217
Nonparametric statistics (**Npar**)
[Normal] (see **Npar/Kolmogorov–Smirnov...**)
Normal CDF (see *Getting Started* and *DATA*)
Normal curve, 325, 447–50

© 1989, SYSTAT, Inc.

Normal distribution, 15, 17, 125, 447
Normal distribution functions, 342
Normal probability plot (see Probability plots)
Normal random numbers (see ZRN)
Normality testing, 17, 359, 369–71
Normalized Euclidean distance, 28, 58, 60
Normalized scores, 86
NOT (see *Getting Started* and *DATA*)
Not equal to (<>) (see *Getting Started* and *DATA*)
Notation (see *Getting Started*)
Npar, 352–72
  Algorithms, 362
  Categorical (nominal) data, 354
  Friedman..., 357–58
    Two-way analysis of variance, 117, 202–06, 238, 365
  Kolmgorov-Smirnov..., 358
    [Chi-square], 370
    [Lilliefors], 359, 370–71
    [Normal], 370
    One-sample, 359, 368–70
    Two-sample, 359, 371–72
    [Uniform], 370
  Kruskal-Wallis..., 353–54, 357, 360, 362, 365–67
    Mann-Whitney U, 353–54, 357, 360
    One-way analysis of variance, 365–67
  Missing data, 362
  Rank (ordinal) data, 354
  Robustness, 355
  Sign test..., 360–61, 363–64
  Ties, 362
  Wald-Wolfowitz runs..., 353, 356, 361, 367–68
    [Cut], 361
  Wilcoxon..., 356, 361–362, 364
    Signed ranks test, 364
NRAN (see ZRN) (see *Getting Started* and *DATA*)
Null hypothesis, 14–15, 125, 168, 177, 201, 218, 293, 356, 457, 475–77
Null hypothesis matrix (D matrix), 135, 144, 153, 168, 188–89, 193, 283, 287–88
[Num of factors] (see **Factor, Nonlin/Model...**)
Numerical values (see *Getting Started* and *DATA*)

© 1989, SYSTAT, Inc.

One-way ANOVA... (see **Stats**)
One-way tables, 472–73, 475–77, 484
Opening a SYSTAT data file (see *Getting Started* and *DATA*)
Operators (see *Getting Started* and *DATA*)
[Options] (see **Factor, Tables/Loglinear models...**)
OR (see *Getting Started* and *DATA*)
Order of expression evaluation (see *Getting Started* and *DATA*)
Ordinal association, 481, 485
Ordinal data, 354
Ordinary least squares, 122–23, 126, 169
Origin, regression through, 172–73
Orthogonal polynomials, 194, 246, 280
Orthogonal rotations, 71
Output device control (see *Getting Started*)
Outside values, 6
Overflow (see *Getting Started*)
Overlapping clusters, 20
Overlapping points, 114

P option, 400, 442
p value, 15, 66–67, 126, 153, 155, 177, 191, 200, 215, 218, 223, 244, 456, 480
PACF... (see **Series**)
Page formatting (see *Getting Started*)
Page titles (see *Getting Started*)
[Paired] (see **Stats/t-test...**)
Pairwise deletion of missing data, 36, 56–57, 84–85, 87, 152, 286, 362
[Pairwise deletion] (see **Factor/Common factor model...**)
[Pairwise] (see **Corr**)
Parameter bounds, 320–21
Parameter intercorrelations, 319–20
Partial autocorrelation function plots (PACF), 150, 188, 389–91, 393, 397–98, 403, 419, 422, 426–430
Pearson... (see **Corr**)
Pearson correlation, 23–25, 28, 38, 51–52, 55, 58, 61–61, 65–67, 96, 159, 385–85
Pearson's chi-square, 480–82, 486–87
[Pearson] (see **Cluster/Join...**)
[Percent change] (see **Series/Transform...**)
[Percent] (see **Tables/Tabulate...**)
[Percentage] (see **Cluster/Join..., Series/Exponential...**)

© 1989, SYSTAT, Inc.

Percentiles, 450
Periodogram, 395–96, 399, 440–41
[Periodicity] (see **Series,ID...**)
Permutation test, 475, 480
Permutation transformations, 354
Phi, 481
Piecewise regression, 333–37
Pillai trace, 243
[Plot factors] (see **Factor**)
[Poly] (see **MGLH/Test...**)
Polynomial regression, 175, 76
Population, definition, 9, 10
Positive skewness, 453
Post hoc pairwise comparisons, 445, 458
Post hoc test, 445, 457
Posterior probabilities (PROB), 291, 294, 299
Power metric, 102
Predictors, 132–33, 143, 150, 159, 160, 163–64, 166, 176–77, 186, 205, 285, 322
[Preview] (see **MGLH/General linear model...**)
Principal component, 72, 74
Principal components analysis, 771–72, 74–75, 79–82, 83–86, 88–90, 139, 155, 164, 190 283, 289–90, 291, 299–304
  Versus factor analysis, 79–82
Principal components... (see **Factor**)
[Print] (see **Nonlin/Model...**)
Printing a SYSTAT data file (see *Getting Started* and *DATA*)
Printing output (see *Getting Started*)
Prior probability, 154, 190, 289, 298–99
[Priors] (see **MGLH/Test...**)
Probabilities for correlation coefficients, 62
Probability plots, 17, 26, 216, 332, 368, 371, 432, 450
[Probability] (see **Corr/Pearson...**)
Probit models, 342–45
Product-moment correlation, 28, 61
Profile analysis, 153–55, 189, 191, 194, 245–50, 254, 278, 280, 288
[Profile] (see **MGLH/Test...**)
Prompt (see *Getting Started* and *DATA*)

QL iterations, 87, 140
Quantile plots, 371

© 1989, SYSTAT, Inc.

Quartiles, 353, 450
[Quartimax] (see **Factor**)
[Quasi-Newton] (see **Nonlin/Model...**)
Quotation marks (see *Getting Started* and *DATA*)

R squared, 159, 228
R-bar squared (see Adjusted squared multiple R)
RAM (see *Getting Started*)
Random distribution functions (see *Getting Started* and *DATA*)
Random distributions (see *Getting Started* and *DATA*)
Random model, 349, 479
Random numbers (see *Getting Started* and *DATA*)
Random sample, 10–11, 13
Random variable, 12, 123
Random variate, 67
Randomized block designs, 219
Range, 120, 447, 450, 453
[Range] (see **Stats/Statistics...**)
Rank data, 355
Rank order coefficients, 64
Rank order data, 354
Rank order transformations, 354
Ranking variables (see *Getting Started* and *DATA*)
Rao's F statistic, 243
[Raw residuals] (see **Tables/Loglinear models...**)
Read external files (see *Getting Started* and *DATA*)
Reading SYSTAT files (see *Getting Started* and *DATA*)
Reading variable names (see *Getting Started* and *DATA*)
Recoding values (see *Getting Started* and *DATA*)
Record length (see *DATA*)
Rectangular SYSTAT files (see *Getting Started* and *DATA*)
Recursive subroutines (see *DATA*)
Regression
  Coefficients, 125, 135, 137, 140, 143–44, 148–49, 152–53, 155–57, 159, 161–64, 165–67, 182, 187, 194, 196, 214, 240, 246, 283, 286
  Constant, 124
  Diagnostics, 126–30
  Function
    Choosing, 102
  Mixture model, 150, 173–75, 186, 285

© 1989, SYSTAT, Inc.

Multiple linear, 130, 143
Polynomial, 175, 76
Simple linear, 120–22
Stepwise, 134, 150, 176–78, 185, 285
Through the origin, 172–73
With ecological data, 168
With grouped data, 168
Without the constant, 188
Regression... (see **MGLH**)
[Regression] (see **MDS/Scale...**)
Relational expression (see *Getting Started* and *DATA*)
Relational operators (see *Getting Started* and *DATA*)
[Remove] (see **MGLH/General linear model...**)
REPEAT...UNTIL loop (see *DATA*)
[Repeat] (see **MGLH/Fully factorial (M)ANOVA...**)
Repeated measures analysis, 150, 154, 184, 186, 254–280
Of covariance, 274–77
One-way, 254–56
With grouping factors, 256–64
With three or more trial factors, 270–74
With two trial factors, 264–270
Replicating cases (see *Getting Started*)
[Rescale MSE] (see **Nonlin/Model...**)
RESIDUAL, 148–49, 160, 164, 187–88, 233
Residual correlation matrix, 178, 293, 301
Residual error, 193
Residual sum of product, 240
Residual sum of squares, 166, 176, 206, 240
Residuals
Definition, 121, 126
Root mean squared, 103
Saving, 146, 148–49, 160, 164–65, 184, 186–88, 287
Studentized, 127
[Residuals] (see **Tables/Loglinear models...**)
Resume... (see **Nonlin**)
Ridge regression, 79
Robustness, 355
Root mean squared distance, 28, 60
Root mean squared residuals, 103
[Rotate] (see **MGLH/Test...**)
Rotated loadings, 85, 303
[Rotation] (see **Factor/Common factor model...**)

© 1989, SYSTAT, Inc.

Round-off error (see *Getting Started* and *DATA*)
[Row pct] (see **Tables/Tabulate...**)
[Rows] (see **Cluster/Join...**)
Roy's union-intersection test, 243
Running means (moving averages), 377–80
Running medians, 380–81
Runs test, 367–68

[S2] (see **Corr/Dichotomous...**)
[S3] (see **Corr/Dichotomous...**)
[S4] (see **Corr/Dichotomous...**)
[S5] (see **Corr/Dichotomous...**)
[S6] (see **Corr/Dichotomous...**)
[Sample] (see **MGLH/General linear model...**)
SAS PROC GLM, 206
Save cluster identifiers, 31
Save estimated distances, 103
[Save file] (see **Corr, Factor, MDS/Scale..., MGLH/General linear model..., Nonlin/Model..., Series/LOWESS..., Series/Fourier..., Series/Seasonal..., Series/Transform...**)
Save final configuration, 102–03
[Save identifiers] (see **Cluster/Join.../Kmeans...**)
[Save residuals] (see **MGLH/Fully factorial (M)ANOVA.../Regression...**)
[Save score and results] (see **MGLH/Test...**)
[Save smoothed values] (see **Series/ARIMA.../Exponential.../Smooth...**)
[Save statistic] (see **Stats/Statistics...**)
[Save table] (see **Tables/Loglinear models...**)
Saving correlation matrix, 63
Saving data in an ASCII file (see *Getting Started* and *DATA*)
Saving results (see *Getting Started*)
Scale... (see **MDS**)
Scatterplot matrix, 25, 50–52
Scatterplot smoothing, 382, 408
Scheffé comparisons, 202
Scheffé confidence intervals, 157
Screen handling capabilities (see *Getting Started*)
Scrolling (see *Getting Started*)
[SD] (see **Stats/Statistics...**)
Searching data (see *Getting Started*)

© 1989, SYSTAT, Inc.

Seasonal... (see **Series**)
[Seasonal AR] (see **Series/ARIMA...**)
Seasonal decomposition and adjustment
[Seasonal MA] (see **Series/ARIMA...**)
Seasonal multiplicative ARIMA model
[Seasonal periodicity] (see **Series/ARIMA.../Exponential.../Seasonal...**)
[Select] (see **Tables/Loglinear models..., Cluster/Kmeans...**)
SEM (see **Stats/Statistics...**)
SEPRED variable, 149, 187
Serially correlated, 128
Seriation, 32
Series, 374–442
ACF plot..., 128, 150, 188, 331, 389–390, 393, 397–398, 403, 419–422, 426–30
[Lag]
Algorithms, 414–15
ARIMA..., 376, 383–93, 397–98, 404, 414–415, 419, 425–33
[AR parameters], 405
[Backcast], 405
[Constant], 405
[Forecast], 405
[MA parameters], 405
[Save forecasts], 405–06]
[Seasonal AR], 405
[Seasonal MA], 405
[Seasonal periodicity], 405
[Tolerance], 405
Autocorrelation, 385–86
Autoregressive models, 386–87
Case plot..., 402
[Lag], 402
[Max], 402
[Min], 402
[Standardize], 402
CCF..., 404, 422–24
[Lag], 404
Clear..., 401

© 1989, SYSTAT, Inc.

Exponential..., 394–95, 408–10, 434–37
[Additive], 408
[Component], 410
[Exponential smoothing], 409
[Forecast], 410
[Linear], 410
[Multiplicative], 410
[Period], 410
[Percentage], 410
[Save forecasts], 410
[Smooth], 409
[Weight], 380–81
Fourier..., 395–400, 412–413, 439–42
[Save file], 412
ID..., 401
[First period], 401
[Periodicity], 401
[Year], 401
LOWESS..., 381–82, 407–08
[Tension], 407
[Save file], 408
Missing data, 414
Moving averages models, 388
PACF..., 150, 188, 389–91, 393, 397–98, 403, 419, 422, 426–430
[Lag], 403
Seasonal..., 393–94, 410–11, 437–38
[Additive], 410–411
[Multiplicative], 411
[Periodicity]
[Save file], 411
Smooth..., 407
[Mean window], 407
[Median window], 407
[Save smoothed values], 407
[Weight], 407
Smoothing
LOWESS, 481–82
Moving averages, 377–80
Running median, 380–81

© 1989, SYSTAT, Inc.

Transform...,398–400
[Base], 398
[Difference], 389, 398, 400
[Index], 398, 399–400
[Lag], 398, 399–400
[Log], 399
[Mean], 399
[Percent change],399
[Save file], 400
[Split-cosine-bell], 398, 400
[Square], 399
[Trend], 389, 393, 399
Shepard diagram, 102
[Shepard plots] (see **MDS/Scale...**)
Shrinkage R-squared (see Adjusted squared multiple R)
Sign test, , 360–61, 363–64
Sign test... (see **Npar**)
Signed ranks test, 356, 361–62, 364
Significance tests, 375–380 (see also *t*–test and *F* test)
Signs of factor loadings, 77
Similarity coefficients, 29, 54, 63
Similarity matrix, 19, 27, 29–30, 58, 93, 100
Simple alternative value, 15
Simple contrasts, 208
Simple data transformations (see *Getting Started* and *DATA*)
[Simplex] (see **Nonlin/Model...**)
Simultaneous confidence interval, 157
SIN function (see *Getting Started* and *DATA*)
Single degree of freedom designs, 212–16
Single linkage clustering (see **Cluster**)
Single precision (see *Getting Started*)
[Single] (see **Cluster/Join...**)
Singular design, 223
Skewness, 453
[Skewness] (see **Stats/Statistics...**)
Smooth... (see **Series**)
[Smooth] (see **Series/Exponential...**)
Smoothing series, (see **Series**/Smoothing)
Somer's D, 495
[Sort matrix] (see **Factor**)
[Sort] (see **Tables/Tabulate...**)
Sorted rotated loadings, 85,

© 1989, SYSTAT, Inc.

Sorting variables (see *Getting Started* and *DATA*)
Sparse cells, 187, 489
Spearman... (see **Corr**)
Spearman rank correlations, 21, 58, 62, 64
Spearman's rho, 52–53, 353, 481
Specificities, 78
Specifying a model, 11–12
Split plot designs, 117, 227–30
Between plots effects, 228
Fit the saturated model, 228
Within plots effects, 229
Split-cosine-bell taper, 398, 400
[Split-cosine-bell] (see **Series/Transform...**)
Spread, 447, 453
SQR function (see *Getting Started* and *DATA*)
Square root function (see *Getting Started* and *DATA*)
[Square] (see **Series/Transform...**)
Squared multiple correlation, 87, 126, 194
Squared multiple R, 159, 228
SSCP... (see **Corr**)
SSCP matrix, 19, 56, 83, 100, 142, 155, 190
As input, 150–52, 285–86
Standard deviation, 5, 447
Standard deviation within groups, 244, 290, 296
[Standardize] (see **Series/Case plot...**)
Standardized coefficients, 132, 144, 151, 159, 286
Standard error, 322, 324, 326, 430, 435, 476
Asymptotic, 319–20, 324, 341, 481
Of estimate, 124, 144, 159
Of predicted values, 149, 187
Of the coefficient, 125, 144, 159, 161, 171
Of the mean, 13–14, 325, 453
Of the parameters, 317, 392, 415
Of predicted values (SEPRED), 149, 187
Of the y-intercept, 124
Printing, 317
Wthin series, 437
Standardizing data, 8–9, 22
[Start values] (see **Nonlin/Model...**)
Starting values, 316, 319, 322–23, 329
Statistics, 2–17, 3
Statistics... (see **Stats**)

© 1989, SYSTAT, Inc.

Stats, 444–469
Algorithms, 459
By groups, 451, 454, 460–61
Missing data, 459
Post hoc pairwise comparisons, 445, 458
Statistics..., 453–55, 460–61
[All], 454
[CV], 454
[Group], 454
[Kurtosis], 453–454
[Maximum], 454
[Mean], 454
[Minimum], 454
[N], 454
[Range], 454
[Save statistic], 454
[SD], 454
[SEM], 454
[Skewness], 454
[Sum], 454
[Variance], 454
Subpopulations, 450–51
t-test..., 445, 451–53, 455–56, 462–64
Dependent (Paired), 456, 463–64
[Independent] (Two–sample), 456, 462–63
Matched pairs, 456
One–sample 457
p value, 456, 480
[Paired] (Dependent), 456, 463–64
Pooled variances, 456, 463
Separate variances, 456, 463
t statistic, 456
One-way ANOVA...
[Tukey HSD], 201, 245, 458–459, 465–69
STD function (see *Getting Started* and *DATA*)
[Stdize] (see **MGLH/Test...**)
Stem-and-leaf displays, 5–6, 353
Stepwise discriminant analysis, 299
Stepwise regression, 134, 150, 176–78, 185, 285
Alpha to enter, 177
Alpha to remove, 178
[Stepwise] (see **MGLH/General linear model...**)

© 1989, SYSTAT, Inc.

Stress, 97–98, 101, 103–04, 106
Structural zeros, 489
Stuart's tau-c, 481
STUDENT, 127, 149, 160, 164, 187
Studentized range, 459
Studentized residuals, 127, 149, 187–88, 209–10
Subgroup means (see *DATA*)
Subgroup processing (see *DATA*)
Subscripted variables (see *Getting Started* and *DATA*)
Subsets of cases (see *DATA*)
Subsets of variables (see *Getting Started* and *DATA*)
Subtraction (–) (see *Getting Started* and *DATA*)
SUM function (see *Getting Started* and *DATA*)
Sum across variables (see *DATA*)
Sum of least squares (see **Loss function...**)
Sum of squares and cross products matrix (see SSCP matrix)
[Sum] (see **MGLH/Test...**, **Stats/Statistics...**)
Summary statistics, 5
Summations (see *Getting Started* and *DATA*)
Summing across cases (see *DATA*)
Sums of squares and cross products matrix (see SSCP matrix)
Suppressor variables, 244
Symmetric confidence interval, 14
Symmetric matrix, 19, 49, 51, 84, 150, 285
SYSTAT BASIC (see *DATA*)
SYSTAT data files (see *Getting Started* and *DATA*)

t distribution, 125, 127, 149, 209, 325
t statistic, 133, 144, 153, 160–61, 171, 210, 456
t-test, 445, 451–53, 455–52, 462–64
Dependent (Paired), 456, 463–64
Independent (Two–sample), 456, 462–63
Matched pairs, 456
One-sample, 457
p value, 456, 480
Paired (Dependent), 456, 463–64
Pooled variances, 456, 463
Separate variances, 456, 463
t statistic, 456
[Table type] (see **Tables/Loglinear models...**)

© 1989, SYSTAT, Inc.

Tables, 470–505
Algorithms, 491
Association measures, 471, 480–81, 483, 486, 494–95
Cross tabulations, 471, 473
Large tables, 486–87, 491
Loglinear models..., 471, 481–83, 487–91, 499–505
[Criterion], 489
[Delta], 489
Diagnostic information, 489
[Fitted], 489
Fully saturated, 481–82, 488
[Iter], 489
[Raw residuals], 489
[Residuals], 489, 499
[Save table], 490
Stuctural zeros, 489
[Zero], 489, 505
Mantel-Haenszel test, 497–98
McNemar's test, 480
Missing values, 486, 491
Multi-way table, 480, 483, 484–85, 500–05
One-way table, 472–73, 475–77, 485
Significance test, 475–480
Standardizing with percents, 473–74
Tabulate..., 297, 483–87, 491–93, 499–502
[Column pct], 485
[Confi], 485
[Frequency], 485
[List format], 485
[Missing], 486
[Percent], 485
[Row pct]485
[Sort], 485
Two-way tables, 473, 477–486, 494–499
Measures of association for, 480–81, 483, 486
McNemar's test, 480
Tabulate... (see **Tables**)
TAN function (see *Getting Started* and *DATA*)
Tangent function (see *Getting Started* and *DATA*)
Tanimoto coefficient, 60
Tau... (see **Corr**)
TCF function (see *Getting Started* and *DATA*)

© 1989, SYSTAT, Inc.

Temporary files (see *DATA*)
[Tension] (see **Series/LOWESS...**), 407
Test... (see **MGLH**)
Test of independence, 353, 471, 479–480, 482
Test of residual roots, 151, 191
Test of simple effects, 181, 193, 208–09
Text editor (see *Getting Started*)
Theta, 243
Ties in nonparametric tests, 362, 364
TIF function (see *Getting Started* and *DATA*)
Time domain models, 376
Time series analysis (**Series...**)
Autocorrelation function plots (see **ACF plot...**)
Autoregressive (AR) modeling, 386–87, 426
Autoregressive moving average (ARMA) modeling, 388
Autoregressive integrated moving average(ARIMA) modeling, 376, 383–93, 397–98, 404, 414–415, 419, 425–33
Backcasting, 405, 430–32
Confirmatory, 376
Cross-correlation function plots (CCF), 404, 419, 422–34
Differencing, 389–92, 399, 406, 414, 416–19, 426–430
Exploratory, 376, 383
Forecasting, 383–93, 405–06, 410, 415, 432–33
Fourier analysis, 376, 393, 395–400, 412–413, 439–42
Moving averages (MA) modeling , 377, 388
Partial autocorrelation plots (PACF), 150, 188, 389–91, 393, 397–98, 403, 419, 422,426–430
Seasonal decomposition, 393–94, 410, 434, 437–38
Smoothing, 376–83, 394, 407–13, 434–37
Exponential, 394–95, 408–410, 434–37
LOWESS, 381–82
Moving averages, 377–80
Running median smoothers, 380–81
Weighted running, 380
Stationarity, 389
Time series smoothing (see Time series analysis/Smoothing)
[Tolerance] (see **Factor/Common factor model...**, **Nonlin/Model...**, **Series/ARIMA...**)
Total sum of product matrix, 291
Transform... (see **Series**)
[Transformations] (see **Series/Transform...**)

© 1989, SYSTAT, Inc.

Transforming data, 398–400, 412–413, 417 (see also *Getting Started* and *DATA*)
Transposing a file (see *Getting Started* and *DATA*)
Tree method of clustering (see **Cluster**)
Tree structure, 21, 32
[Trend] (see **Series/Transform...**)
Triads, 95
Trials factors, 259, 264–74
Trials interaction, 269, 272–273, 278
Triangle inequality, 94–95
Triangular matrices, 152, 286, 496
Trigonometric functions (see *Getting started* and *DATA*)
TRN function (see *Getting Started* and *DATA*)
Tukey HSD test, 201, 245, 458–59, 465–69
[Tukey HSD test] (see **Stats/One-way ANOVA...**)
Tukey's gapping method, 33
Tukey–Kramer adjustment, 458
Tutorial (see *Getting Started*)
Two-way analysis of variance, 117, 202–10, 238, 365
  Post-hoc tests, 206
  Screening results, 209, 210
  Test of simple effects, 181, 191, 208–09
Two-way clustering (see **Cluster**)
Two-way table (see **Tables**)
Type III sum of squares, 206, 218

U statistic (Mann-Whitney U), 353–54, 357, 360
UCF function (see *Getting Started* and *DATA*)
UIF function (see *Getting Started* and *DATA*)
Unary minus function (see *Getting Started* and *DATA*)
Unbalanced designs, 181, 187, 206, 216–18, 227, 277
Uncertainty, 495
Unevenly spaced polynomials, 247–48
Uniform distribution functions (see *Getting Started* and *DATA*)
Uniform integers (see *Getting Started* and *DATA*)
Uniform random numbers (see *Getting Started* and *DATA*)
[Uniform] (see **Npar/Kolmogorov-Smirnov...**)
Union-intersection test, 243
Unweighting means, 250–54
URN function (see *Getting Started* and *DATA*)

© 1989, SYSTAT, Inc.

Variable length records (see *Getting Started* and *DATA*)
Variables (see *Getting Started* and *DATA*)
Variance, Analysis of (see Analysis of variance)
Variance conservation, 76
[Variance] (see **Stats/Statistics...**)
Variance proportions, 157, 164
Variates, 67, 244
[Varimax] (see **Factor/Rotations**)
Viewing a SYSTAT data file, (see *Getting Started*)
Violations of assumptions, 197, 211

Wald-Wolfowitz runs... (see **Npar**)
Ward's method of clustering (see **Cluster**)
[Ward] (see **Cluster/Join...**)
Warning messages (see *Getting Started*)
[Weight] (see **Series/Exponential.../Smooth...**)
Weighted estimates, 170–72, 327
Weighted least squares, 143, 150, 168–71, 380
Weighted running, 380
Wherry R squared (see Adjusted squared multiple R)
WHILE loop (see *DATA*)
White noise, 390, 392
Wilcoxon... (see **Npar**)
Wilcoxon signed ranks test, 356, 361–62, 364
Wilks' lambda, 243,299
Windows (see *Getting Started*)
Winter's three-parameter model, 395, 409
[Within groups], 244
Within groups covariance matrix, 154, 190
Within groups principal components analysis, 289, 299–304
Within groups standard deviations, 244, 290, 296
Within groups sum of squares, 25, 278
Within subjects effects, 258–67, 274, 278–79
Writing SYSTAT files (see *Getting Started* and *DATA*)

XCF function (see *Getting Started* and *DATA*)
XIF function (see *Getting Started* and *DATA*)
XRN function (see *Getting Started* and *DATA*)

© 1989, SYSTAT, Inc.

Yates' chi-square test, 480
[Year] (see **Series/ID...**)
Yule's Q and Y, 481

ZCF function (see *Getting Started* and *DATA*)
Zero-intercept models, 489, 505
[Zero] (see **Tables/Loglinear models...**)
ZIF function (see *Getting Started* and *DATA*)
ZRN function (see *Getting Started* and *DATA*)

© 1989, SYSTAT, Inc.